Genomics driven metabolomics
-
Novel strategies for the discovery and identification of secondary metabolites

Genomics driven metabolomics - Novel strategies for the discovery and identification of secondary metabolites
Thesis, Leiden University, Leiden

ISBN: 9789074538817

Cover illustration: Secondary metabolites produced by *Penicillium chrysogenum* drawn as balls-and-sticks.Biosynthetic reactions, performed by the fungus, are indicated by dashed arrows. High structural similarity between two compounds, based on fragmentation tree comparison, is represented by solid two-headed arrows

Printed by Ipskamp Drukkers B.V.

Genomics driven metabolomics
-
Novel strategies for the discovery and identification of secondary metabolites

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 25 februari 2014
klokke 11.15 uur

door
Marco Ries

geboren te Lauchhammer, Duitsland

in 1982

PROMOTIECOMMISSIE

Promotor:	prof. dr. T. Hankemeier
Co-promotor:	dr. R. J. Vreeken
Overige leden:	prof. dr. R. A. L. Bovenberg University of Groningen, the Netherlands
	prof. dr. M. Danhof Leiden University, the Netherlands
	prof. dr. A. J. M. Driessen University of Groningen, the Netherlands
	prof. dr. R. Goodacre University of Manchester, United Kingdom
	prof. dr. G. P. van Wezel Leiden University, the Netherlands

This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO) and partly funded by the Ministry of Economic Affairs (project number 10469).

TABLE OF CONTENTS

Chapter 1

General Introduction

General introduction

Metabolic Pathway Analysis: Basic concepts and scientific applications in the post-genomic era

Natural products have been of major interest due to their pharmaceutically exploitable properties like antibacterial (Christophersen, et al., 1998), antifungal (Coleman, et al., 2011), antiparasitic (Wink, 2012), anticancer (Nirmala, et al., 2011) or immunosuppressive (von Wartburg and Traber, 1988) activities. Although not always used directly as a drug they act as a reliable source for novel and innovative therapeutic agents offering valuable chemical scaffolds for novel drugs (Newman and Cragg, 2007). The primary source of these structurally heterogenic molecules are natural sources like bacteria and fungi, which is not surprising as these microorganisms live in complex ecosystems where they compete and communicate with other organisms (Goh, et al., 2002; Losada, et al., 2009) Up till today, natural products represent the largest source of drugs offering novel chemical entities. Although newer techniques like combinatorial chemistry have been used as a discovery source for several years, solely one approved drug with a new chemical entity was reported to be discovered between 1982 and 2007 using this method (Newman and Cragg, 2007). In 2005, Bayer's antitumor compound sorafenib, obtained from combinatorial chemistry, was approved by the FDA. This gives reason to expect that microorganisms will play also a major role in future drug developments.

However, the incredible increase in the amount of sequencing data, primarily driven by the falling cost of DNA sequencing, has led to a paradigm shift in the approach of secondary metabolite discovery (Letzel, et al., 2013). The analysis of the growing number of sequenced fungal genomes revealed that most of the genes responsible for secondary metabolite production are located in clusters which far outweigh the number of described secondary metabolites (Brakhage and Schroeckh, 2011). That means, that despite the discovery of countless natural products during the last decades a vast number of novel natural products with potent bioactivities based on novel structural features still awaits discovery. With the use of genome sequencing data unknown secondary metabolite clusters can be identified without a priori knowledge of a strain's ability to produce natural products. This approach of secondary metabolite discovery is known as genome mining and has been successfully applied for the identification of various novel metabolites (Letzel, et al., 2013).

These metabolites are produced by large, multifunctional protein complexes called non ribosomal peptide synthetases (NRPS) or polyketide synthetases (PKS) which catalyze the stepwise condensation of simple amino acids or malonyl building blocks to complex molecules. Although their substrates can differ considerably, these multi-modular enzymes show striking similarities in their architecture as well as in the mechanisms used for product assembly.

NRPSs have a modular organization, with each module responsible for one or more chain-elongation step. Every single module is subdivided into three basic domains that carry all essential information for recognition, activation and modification of one substrate. At a minimum, a typical NRPS module consists of an adenylation (A) domain responsible for amino acid activation, a thiolation (T) domain, also

known as peptidyl carrier protein (PCP), which binds the activated amino acid and a condensation (C) domain that catalyzes peptide-bond formation. The common arrangements of these domains follow a $(C\text{-}A\text{-}PCP)_n$ organization. Additionally, a variety of optional domains have been described (Schwarzer, et al., 2003) such as methyltransferase (MT) and epimerization (E) domains (Figure 1).

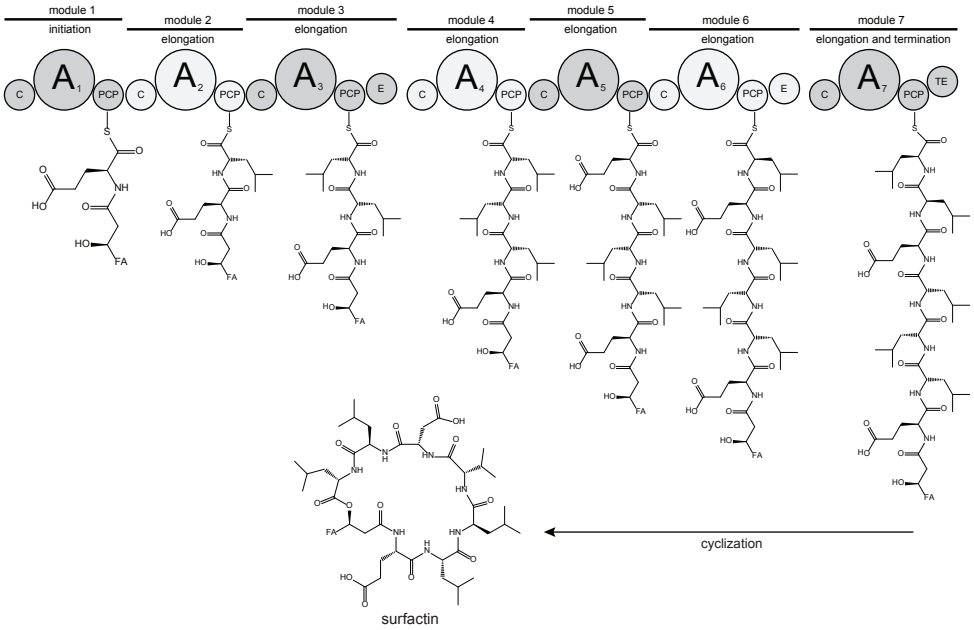


Figure 1. NRPS assembly of surfactin
The surfactin biosynthesis consists of seven modules each responsible for one chain-elonging step. Each module is divided in various domains catalyzing the activation (A domain), covalent binding (PCP domain), elongation (C domain), epimerization (E domain) and termination by covalent cyclization (TE domain).

A similar domain architecture can be found in PKS systems with an acyltransferase (AT) domain responsible for building block selection and transfer, an acyl carrier protein (ACP) for unit loading and a ketosynthase (KS) domain for decarboxylative condensation of the extending substrate. The resulting keto thioester can subsequently be modified by ketoreductase (KR) domains, dehydratase (DH) domains and enoyl reductase (ER) domains (Cane, 1997; Staunton and Weissman, 2001). In addition, modular NRPS and PKS systems can closely cooperate to form a third group of gene clusters, so-called hybrid products.
The number of modules and their domain organization within the enzymes control the structures of the final products (Grunewald and Marahiel, 2006; Schwarzer, et al., 2003; Schwarzer, et al., 2002). Thus, the order of modules usually corresponds to the sequence of building blocks in the final product. Many systems adhere to this mechanistic paradigm, which is often referred to as the “co-linearity rule” (Fischbach and Walsh, 2006). However, several exceptions to this rule have been discovered in

the last years including iterative NRPS and PKS which incorporate multiple residues of the same unit iteratively into the final structure and the so called nonlinear NRPSs which deviate completely from the standard domain organization leading to unexpected products (Mootz, et al., 2002; Shaw-Reid, et al., 1999; Shen, 2003). In 2008, the genome of Penicillium chrysogenum Wisconsin54-1255 was sequenced revealing 20 polyketide synthases (PKS), 10 nonribosomal peptide synthetases (NRPS) and 2 hybrid NRPS-PKS genes clusters (van den Berg, 2008). According to the deduced domain organization up to 20 polyketides, 10 nonribosomal peptides and two products with mixed properties are expected. However, most gene clusters encoding biosynthetic systems could not be associated with the production of a known metabolite and are therefore referred to as 'cryptic' or 'orphan'. The reason for the lack of assignment is that under standard laboratory conditions the majority of secondary metabolite biosynthesis gene clusters is not expressed (van den Berg, 2011). These clusters remain silent as long as the triggers for their induction have not been identified. (Brakhage, et al., 2008). Furthermore, due to the complexity of these clusters almost no products could be directly deciphered from the genetic sequence with sufficient accuracy and confidence (van den Berg, 2011).

Strategies for the discovery of new natural products

Several strategies can be applied to elucidate the function of biosynthetic gene clusters, depending on the expression of latter (Challis, 2008). For expressed cryptic genes inactivation of a biosynthetic gene, which is presumed to be essential within the biosynthetic gene cluster, followed by comparative metabolite profiling of the wild-type organism and the non-producing mutant is the standard approach for defining the functions of genes and for the discovery of novel natural products (Brakhage and Schroeckh, 2011) (Figure 2). The metabolites present in the wild-type but missing in the mutant are likely to be products of the cryptic gene cluster and can be isolated and structurally characterized. (Figure 2) Alternative strategies are applied when the gene cluster of interest is silent involving promoter exchange, overexpression of transcription factors or other pleiotropic regulators. (Brakhage and Schroeckh, 2011; Chiang, et al., 2011; Schumann and Hertweck, 2006)

All approaches combine an elaborate and refined strategy to discover, isolate, and characterize novel natural products. They all have in common that sensitive, accurate, quantitative and fast analytical techniques are required to detect differences in production related to the conducted genetic modification. Furthermore, due to the rapid advances in DNA sequencing technology it is conceivable that, in the recent future, thousands of cryptic natural product biosynthetic gene clusters will become available, leaving the discovery of the metabolic products of these clusters the bottleneck (Zerikly and Challis, 2009).

Metabolite discovery and metabolomics

The aim of metabolomics is to measure the full metabolome, which consists of all low molecular weight species in cells, tissues, organs or organisms (Griffin, 2004). Unlike the proteome, which represents the entire set of proteins, primarily build from 20 basic amino acids with a limited set of modifications and a rather defined distribution of physicochemical properties, are metabolites a highly diverse range of compounds differing in mass, size and polarity. With concentrations ranging

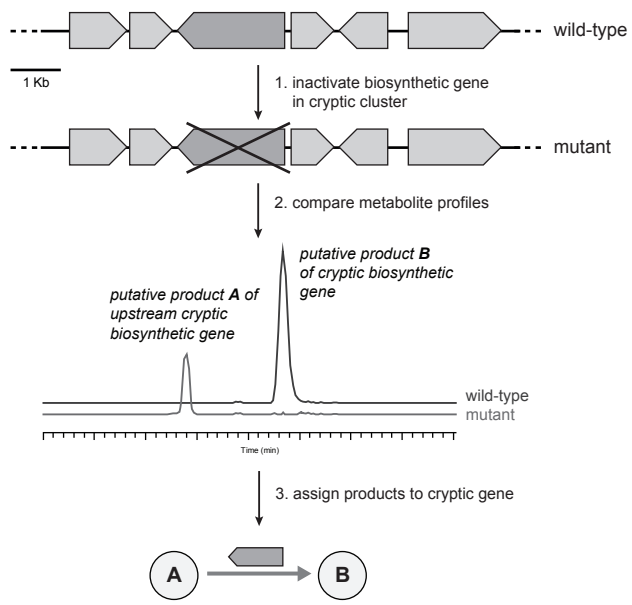


Figure 2. Principle of a gene knockout strategy in combination with comparative metabolite profiling to elucidate the role of a cryptic biosynthetic gene.

over several orders of magnitude (pM – mM) combined with often a short lifetime represent metabolites a challenging class of compounds to analyze. Different from transcriptomics and genomics, which utilize polymerase chain reaction (PCR) to amplify DNA sequences, there is no effective tool to unbiasedly increase the concentration of all low abundant metabolites. Therefore, more than one analytical method is required for the acquisition of a comprehensive metabolite profile attempting to describe the global metabolism (van der Greef, et al., 2007). Even today, despite the rapid advances in technical instrumentation in the last years, no analytical tool is capable to completely cover the entire metabolome of even the simplest organism (Heather, et al., 2013).

The two most commonly used techniques in metabolomics are NMR spectroscopy and mass spectrometry. NMR is based on the measurement of isotopes that contain an odd number of protons and/or neutrons in a magnetic field. Typically, metabolites higher concentrated than 5-10 µM can be detected if they are not co-resonant with higher concentrated metabolites. (Heather, et al., 2013). Despite this limited sensitivity (Martin, et al., 2007), NMR spectroscopy is widely used in metabolomics studies due to its speed, non-invasiveness and robustness and its highly quantitative nature. As the area of a peak is proportional to the concentration of the metabolite measured, determination of absolute concentrations is possible.

In contrast, mass spectrometry (MS) techniques which measure molecules as ions in the gas phase after ionization, are fast and highly sensitive with a large dynamic range. By fragmenting these ions, structural information can be obtained supporting the identification of relevant metabolites (Dettmer, et al., 2007). As a wide range of different MS instruments are available, ranging from high resolution MS such as Fourier transform ion cyclotron resonance (FTICR), Orbitrap FT and time-of-light

(TOF) to low resolution MS (ion traps, triple and single quads) up to hybrid systems, various techniques with different emphasis can be applied.
However, mass spectrometry is most commonly coupled to chromatographic separation techniques like gas chromatography (GC-MS) and liquid chromatography (LC-MS) allowing the separation of complex mixtures into its individual components. By doing so, LC-MS and GC-MS exceed the sensitivity of NMR spectroscopy by many orders of magnitude. For these reasons, mass spectrometry is the primarily used technique in metabolomics studies currently outnumbering the application of NMR (Dettmer, et al., 2007).
For GC–MS, metabolites have to be volatile or made volatile using derivatization techniques. Especially small molecules like amino acids, sugars, glycolytic intermediates, fatty acids and TCA cycle intermediates can be detected after derivatization using derivatization reagents like N-methyltrimethylsilyltrifluoroacetamide (MSTFA). The most common ionization technique in GC-MS is electron ionization (EI) which produces, next to the radical molecule cation, characteristic fragments for many metabolites allowing their identification with databases like NIST or the Golm Metabolome Database (Kopka, et al., 2005). While a robust and versatile technique, GC–MS is limited to compounds which are or can be made volatile and thermostable enough for application in the system. In cases where this cannot be achieved, LC–MS is an option which in principle does not require derivatization, allowing fast analysis.
Depending on the polarity of the analytes of interest, a suitable separation technique like normal-phase liquid chromatography (NPLC), reversed-phase liquid chromatography (RPLC) or hydrophilic interaction chromatography (HILIC) (Tolstikov and Fiehn, 2002) can be chosen for an efficient separation of the individual compounds. After chromatographic separation, the analytes are introduced into the mass spectrometer using an ionization technique like electrospray ionization (ESI) or less frequent atmospheric pressure chemical ionization (APCI) (Niessen, 2003). As minimal fragmentation takes place during ionization, the measured mass of an analyte, recorded with high mass accuracy and resolution, is expected to be close to the anticipated mass of a specific metabolite recorded in a publically accessible database which potentially allows a direct putative identification. Several databases containing metabolite information are available in order to facilitate putative annotation of accurate mass signals. (e.g. PubChem; HMDB; KEGG, KNapSAcK, MZedDB).

Metabolite Profiling
Metabolomics experiments can be conducted as non-targeted profiling using an unbiased approach which aims to measure as many metabolites as possible. As the conditions are not optimized for a particular set of compounds, limits of detection can be severely impaired resulting, at worst, in a failure to detect and identify the desired metabolites.
If more information about the metabolites of interest is available, more targeted profiling approaches can be used focusing on a particular class of compounds. As the biosynthesis of secondary metabolites by PKS or NRPS systems underlays modular based mechanisms, bioinformatics tools can be used to predict substrates, their subsequent reactions and potential products (Figure 3). The calculation of their

corresponding physiochemical properties facilitates the subsequent discovery of these natural products from biological origins as an appropriate targeted profiling method, optimized for their detection, can be selected. Often a non-targeted profiling approach yielding semi-quantitative data is followed by a targeted profiling

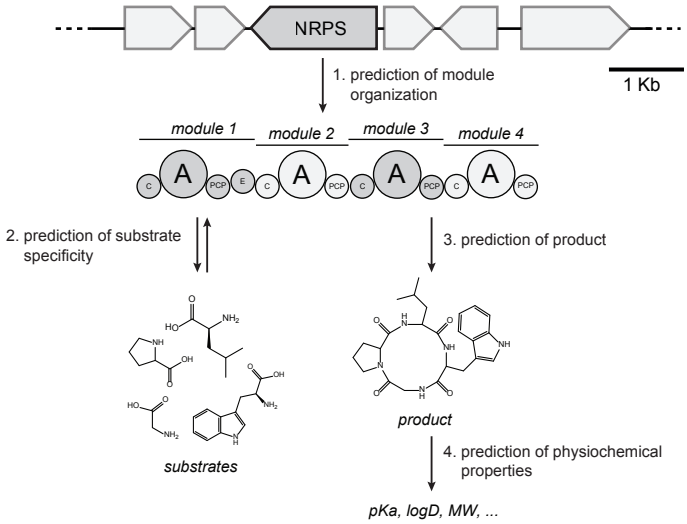


Figure 3. Bioinformatics driven selection of an analytical approach for metabolite discovery
Prediction of the modular organization of a cryptic PKS/NRPS system in combination with calculated substrate specificities results in possible products synthesized by the biosynthetic gene cluster. Predicting their physiochemical properties can help to select appropriate analytical techniques for their detection.

method which allows the determination of absolute concentrations. By identifying metabolites varying between wild-type and mutant samples using statistical methods like the Student's t-test, analysis of variance (ANOVA) or principal component analysis (PCA), metabolic changes can be identified which are related to the genetic modification. This approach, called comparative metabolite profiling, is the most often used approach for the determination of metabolic differences (Fernie, et al., 2004).

Structure elucidation in metabolomics

Metabolite identification is an essential part of any metabolomics study and frequently cited as the major bottleneck in metabolomics (Scalbert, et al., 2009). Without a proper identification, the discovery of metabolic alterations related to a biological question is uninterpretable and limited to diagnostic purposes only. However, as one of the most challenging and labor intensive parts in the metabolomics framework, metabolite identification is neglected in many studies or even completely ignored as the distinctive versatility of metabolites, with millions of possible structures even matching a given elemental composition, imposes a major challenge.

Commonly, complementary analytical techniques like NMR, IR, UV, MS, etc. are employed with each demanding different properties from a given sample. Mass

spectrometry in combination with various fragmentation techniques represents one of the most routinely used techniques for metabolite identification as it requires small sample volumes, low sample concentrations and purity and short analysis times. By fragmenting ionized molecules and detect their charged fragments structural information can be obtained. Using this approach in combination with the identified elemental composition of a metabolite, MS/MS databases can be built and screened in order to match an identical database entry (Horai, et al., 2010; Wishart, et al., 2009). However, different levels of identification are defined depending on the available metadata (Sumner, et al., 2007). Whereas on level 1 two independent and orthogonal techniques are required for a full identification of a non-novel metabolite using a reference compound, represents a level 2 identification a putatively annotated compound without chemical reference standards available but identity to metadata from an entry in a database. A level 3 identification is the putatively characterization of a compound based on similarities to a specific class of compounds. Metabolites identified at level 4 represent unknown compounds which can be differentiated and quantified but are structurally not characterized.

Scope of the thesis

The aim of this thesis is to demonstrate how to discover, identify and ultimately assign novel secondary metabolites with new structural features harboring potentially pharmaceutically exploitable properties to their corresponding NRPS and/or PKS gene clusters in Penicillium chrysogenum using genome mining strategies. For this, an adequate analytical pipeline is required which allows the unbiased detection of genetic modifications on a metabolic level. As products and intermediates of secondary metabolite gene clusters show complex chemical structures and are mostly present only at low concentrations, an advanced mass spectrometry based pipeline was developed which enables the identification of structural complex metabolites at low concentrations directly from a biological matrix.

Outline of the thesis

In Chapter 2, the validation and further application of a targeted profiling method is described for the identification of secondary metabolites originating from a cryptic di-modulated NRPS gene cluster. As bioinformatics analysis of the gene cluster predicted various modified dipeptides involved in the biosynthetic pathway, a reversed phase based separation method was developed. After structure elucidation of metabolites using NMR and MS/MS based techniques, detailed enzymatic reaction steps were identified and their corresponding genes assigned.

In Chapter 3, the identification of structurally novel compounds originating from the roquefortine/meleagrin pathway is described and their impact on the previously proposed pathway discussed. As a results, new biosynthetic reactions were found, indicating excessive branching of the previously as linear reported pathway, with several pharmaceutically interesting end products.

The application of the secondary metabolite profiling pipeline for samples derived from genetic modifications of a cryptic tetra-modular NRPS system is subject of Chapter 4. As this non-linear NRPS lacks distinctive substrate specificity, various novel tetrapeptides could be found, harboring similar physiochemical properties,

thus making their discovery and identification challenging. In combination with substrate predictions, a detailed biosynthetic mechanism is proposed.
In Chapter 5, a novel mass spectrometry based structure elucidation pipeline for the de novo structure identification of small molecules, coined CMCP (Chemoinformatics supported MSn Comparison Pipeline) is presented. By transferring fragmentation mechanisms from a similar fragmenting database entry, complete structures could be identified solely using mass spectrometry. To demonstrate the various concepts of fragmentation tree comparison, the de novo identification of structurally complex secondary metabolites, obtained from comparative metabolite profiling from Chapter 2 and 3, is described.
In Chapter 6, the application of CMCP in direct infusion based experiments is demonstrated. Challenges and advantages resulting from the lack of a separation step prior MS analysis are discussed. To demonstrate the capabilities of this approach to cope even with co-fragmenting interferences like isobaric and isomeric ions, the structure elucidation of various complex metabolites from liquid cultures of Penicillium chrysogenum is described using DI-MS in combination with CMCP.

References

Brakhage, A.A., and Schroeckh, V. (2011). Fungal secondary metabolites - strategies to activate silent gene clusters. Fungal Genet Biol 48, 15-22.

Brakhage, A.A., Schuemann, J., Bergmann, S., Scherlach, K., Schroeckh, V., and Hertweck, C. (2008). Activation of fungal silent gene clusters: a new avenue to drug discovery. Prog Drug Res 66, 1, 3-12.

Cane, D.E. (1997). Introduction: Polyketide and Nonribosomal Polypeptide Biosynthesis. From Collie to Coli. Chem Rev 97, 2463-2464.

Challis, G.L. (2008). Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology 154, 1555-1569.

Chiang, Y.M., Chang, S.L., Oakley, B.R., and Wang, C.C. (2011). Recent advances in awakening silent biosynthetic gene clusters and linking orphan clusters to natural products in microorganisms. Curr Opin Chem Biol 15, 137-143.

Christophersen, C., Crescente, O., Frisvad, J.C., Gram, L., Nielsen, J., Nielsen, P.H., and Rahbaek, L. (1998). Antibacterial activity of marine-derived fungi. Mycopathologia 143, 135-138.

Coleman, J.J., Ghosh, S., Okoli, I., and Mylonakis, E. (2011). Antifungal activity of microbial secondary metabolites. PLoS One 6, e25321.

Dettmer, K., Aronov, P.A., and Hammock, B.D. (2007). Mass spectrometry-based metabolomics. Mass Spectrom Rev 26, 51-78.

Fernie, A.R., Trethewey, R.N., Krotzky, A.J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5, 763-769.

Fischbach, M.A., and Walsh, C.T. (2006). Biochemistry. Directing biosynthesis. Science 314, 603-605.

Goh, E.B., Yim, G., Tsui, W., McClure, J., Surette, M.G., and Davies, J. (2002). Transcriptional modulation of bacterial gene expression by subinhibitory concentrations of antibiotics. Proc Natl Acad Sci U S A 99, 17025-17030.

Griffin, J.L. (2004). Metabolic profiles to define the genome: can we hear the phenotypes? Philos Trans R Soc Lond

B Biol Sci 359, 857-871.

Grunewald, J., and Marahiel, M.A. (2006). Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. Microbiol Mol Biol Rev 70, 121-146.

Heather, L.C., Wang, X., West, J.A., and Griffin, J.L. (2013). A practical guide to metabolomic profiling as a discovery tool for human heart disease. J Mol Cell Cardiol 55, 2-11.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45, 703-714.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., et al. (2005). GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21, 1635-1638.

Letzel, A.C., Pidot, S.J., and Hertweck, C. (2013). A genomic approach to the cryptic secondary metabolome of the anaerobic world. Nat Prod Rep 30, 392-428.

Losada, L., Ajayi, O., Frisvad, J.C., Yu, J., and Nierman, W.C. (2009). Effect of competition on the production and activity of secondary metabolites in Aspergillus species. Med Mycol 47 Suppl 1, S88-96.

Martin, F.P., Dumas, M.E., Wang, Y., Legido-Quigley, C., Yap, I.K., Tang, H., Zirah, S., Murphy, G.M., Cloarec, O., Lindon, J.C., et al. (2007). A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model. Mol Syst Biol 3, 112.

Mootz, H.D., Schwarzer, D., and Marahiel, M.A. (2002). Ways of assembling complex natural products on modular nonribosomal peptide synthetases. Chembiochem 3, 490-504.

Newman, D.J., and Cragg, G.M. (2007). Natural products as sources of new drugs over the last 25 years. J Nat Prod 70, 461-477.

Niessen, W.M. (2003). Progress in liquid chromatography-mass spectrometry instrumentation and its impact on high-throughput screening. J Chromatogr A 1000, 413-436.

Nirmala, M.J., Samundeeswari, A., and Sankar, P.D. (2011). Natural plant resources in anti-cancer therapy-A review. Research in Plant Biology 1, 1-14.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B.S., van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D., and Wopereis, S. (2009). Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics 5, 435-458.

Schumann, J., and Hertweck, C. (2006). Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. J Biotechnol 124, 690-703.

Schwarzer, D., Finking, R., and Marahiel, M.A. (2003). Nonribosomal peptides: from genes to products. Nat Prod Rep 20, 275-287.

Schwarzer, D., Mootz, H.D., Linne, U., and Marahiel, M.A. (2002). Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. Proc Natl Acad Sci U S A 99, 14083-14088.

Shaw-Reid, C.A., Kelleher, N.L., Losey, H.C., Gehring, A.M., Berg, C., and Walsh, C.T. (1999). Assembly line enzymology by multimodular nonribosomal peptide synthetases: the thioesterase domain of E. coli EntF catalyzes both elongation and cyclolactonization. Chemistry & biology 6, 385-400.

Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr Opin Chem Biol 7, 285-295.

Staunton, J., and Weissman, K.J. (2001). Polyketide biosynthesis: a millennium review. Nat Prod Rep 18, 380-416.

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., et al. (2007). Proposed minimum reporting standards for chemical analysis. Metabolomics 3, 211-221.

Tolstikov, V.V., and Fiehn, O. (2002). Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. Anal Biochem 301, 298-307.

van den Berg, M.A. (2011). Impact of the Penicillium chrysogenum genome on industrial production of metabolites. Appl Microbiol Biotechnol 92, 45-53.

van den Berg, M.A.e.a. (2008). Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. Nature biotechnology 26, 1161-1168.

van der Greef, J., Martin, S., Juhasz, P., Adourian, A., Plasterer, T., Verheij, E.R., and McBurney, R.N. (2007). The art and practice of systems biology in medicine: mapping patterns of relationships. J Proteome Res 6, 1540-1559.

von Wartburg, A., and Traber, R. (1988). Cyclosporins, fungal metabolites with immunosuppressive activities. Prog Med Chem 25, 1-33.

Wink, M. (2012). Medicinal plants: a source of anti-parasitic secondary metabolites. Molecules 17, 12771-12791.

Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., et al. (2009). HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37, D603-610.

Zerikly, M., and Challis, G.L. (2009). Strategies for the discovery of new natural products by genome mining. Chembiochem 10, 625-633.

Chapter 2

A branched biosynthetic pathway is involved in production of roquefortine and related compounds in *Penicillium chrysogenum*

Based on

Hazrat Ali*, Marco I. Ries*, Jeroen G. Nijland, Peter P. Lankhorst, Thomas Hankemeier, Roel A.L. Bovenberg, Rob J. Vreeken, Arnold J.M. Driessen
A branched biosynthetic pathway is involved in production of roquefortine and related compounds in Penicillium chrysogenum
PLos ONE, **2013***, 8 (6), e65328*

** these authors contributed equally*

Abstract

Profiling and structural elucidation of secondary metabolites produced by the filamentous fungus *Penicillium chrysogenum* and derived deletion strains were used to identify the various metabolites and enzymatic steps belonging to the roquefortine/meleagrin pathway. Major abundant metabolites of this pathway were identified as histidyltryptophanyldiketopiperazine (HTD), dehydrohistidyltryptophanyldiketopiperazine (DHTD), roquefortine D, roquefortine C, glandicoline A, glandicoline B and meleagrin. Specific genes could be assigned to each enzymatic reaction step. The nonribosomal peptide synthetase RoqA accepts L-histidine and L-tryptophan as substrates leading to the production of the diketopiperazine HTD. DHTD, previously suggested to be a degradation product of roquefortine C, was found to be derived from HTD involving the cytochrome P450 oxidoreductase RoqR. The dimethylallyltryptophan synthetase RoqD prenylates both HTD and DHTD yielding directly the products roquefortine D and roquefortine C without the synthesis of a previously suggested intermediate and the involvement of RoqM. This leads to a branch in the otherwise linear pathway. Roquefortine C is subsequently converted into glandicoline B with glandicoline A as intermediates, involving two monooxygenases (RoqM and RoqO) which were mixed up in an earlier attempt to elucidate the biosynthetic pathway. Eventually, meleagrin is produced from glandicoline B involving a methyltransferase (RoqN). It is concluded that roquefortine C and meleagrin are derived from a branched biosynthetic pathway.

Introduction

Fungi produce a variety of secondary metabolites, which have diverse activities, ranging from natural antibiotics to simple toxins or immunosuppressants (Demain and Fang, 2000). Most of these metabolites are synthesized either by nonribosomal peptide synthetases (NRPS) or polyketide synthases (PKS) and can be further modified by a range of biosynthetic enzymes. The genes encoding the NRPS or PKS and the modifying enzymes are usually located in the same genetic region and are often co-expressed (Keller, et al., 2005; Shwab and Keller, 2008). Several methods have been developed to identify the secondary metabolite biosynthetic gene clusters and their respective products. Key methods are heterologous gene expression, bioinformatic (SMURF) identification of gene clusters (Khaldi, et al., 2010) and genome mining strategies in which core synthetase genes are deleted and changes in metabolite production are identified by comparative metabolite profiling (Challis, 2008). Secondary metabolites genes and/or their clusters were frequently transferred from one organism to another during evolution (Donadio, et al., 2005; Fischbach, et al., 2008). The distribution of these clusters in fungal genomes supports the hypothesis that different organisms produce similar metabolites (Ma, et al., 2010). Various fungal species were reported to produce diketopiperazines, a class of pharmaceutically important naturally produced secondary metabolites (Maristela and Carvalho, 2007; Scherlach and Hertweck, 2006). Roquefortine C, a diketopiperazine that was first isolated from *Penicillium roquefortine* (Scott and Kennedy, 1976) has now been reported from 25 different *Penicillium* species (Frisvad, et al., 2004). It displays bacteriostatic activity against Gram-positive bacteria (Kopp-Holtwiesche and Rehm, 1990). Although the exact mechanism of action is not known it appears to interact with cytochrome p450 and interferes with RNA synthesis (Aninat, et al., 2001; Kopp and Rehm, 1981). Roquefortine C also shows neurotoxic activity in mice and is considered as a contaminant in blue cheese (Polonsky, et al., 1977; Scott, et al., 1979). Meleagrin, a downstream product of roquefortine C, has been proposed to be the precursor of neoxaline, a compound with antimicrobial activity (Overy, et al., 2005).

The filamentous fungus *P. chrysogenum* has been explored for more than eighty years for its excellent fermentation capacity and penicillin production (Fleming, 2001; Weber, et al., 2012). In recent years, genome sequencing, as well as microarray analysis in combination with genetic modification of *P. chrysogenum*, has provided a strong base to elucidate the role of the secondary metabolite clusters found in this organism. The genome of *P. chrysogenum* encodes 20 putative PKS and 11 NRPS genes including the gene encoding δ-(L-α-aminoadipyl)-L-cysteinyl-D-valine synthetase involved in penicillin biosynthesis (van den Berg, et al., 2008). This allowed us to identify the corresponding biosynthetic genes of the roquefortine/meleagrin pathway using a gene deletion strategy as well as quantitative metabolite profiling. Here, we revisit an earlier proposal for a linear pathway for the biosynthesis of roquefortine C that included several ambiguities as well speculations because of incomplete metabolite profiling and enzyme miss assignment. We have elucidated the entire biosynthetic pathway up to meleagrin production including a series of precursors. It is demonstrated that meleagrin is derived from a branched pathway with two biologically derived diketopiperazines as major intermediates.

Experimental procedures

Chemicals
HPLC-grade acetonitrile and methanol were purchased from Biosolve (the Netherlands). Dichloromethane, formic acid the internal standards reserpine, ranitidine and ampicillin were acquired from Sigma-Aldrich (St. Louis, MO). Meleagrin and neoxaline were purchased from Bio-Connect (the Netherlands). Roquefortine C was obtained from Bioaustralis (Australia). L-Tryptophan, L-Histidine and Mevalonic acid Lactone were purchased from Sigma-Aldrich. All purchased compounds were of highest available purity.Animal studies

Host strains, media, grown condition and plasmid construction
P. chrysogenum strain DS54555, which lacks penicillin cluster genes and *Ku70* gene was used as a host strain for deletion analysis and was kindly supplied by DSM Anti-infective. All the strains were grown on YGG-medium (Kovalchuk, et al., 2012) for protoplasts formation and transformation. For analysis, cells were grown on SMP medium (glucose, 5.0 g/L; lactose, 75 g/L; urea, 4.0 g/L; Na_2SO_4, 4.0 g/L; CH_3COONH_4, 5.0 g/L; K_2HPO_4, 2.12 g/L; KH_2PO_4, 5.1 g/L) for secondary metabolites production using a shaking incubator at 200 rpm for 168 hours at 25°C. Deletion plasmids were constructed by amplifying the flanking regions of the targeted gene with the Multisite Gateway®Three-Fragment Vector Construction Kit (Invitrogen). *Escherichia coli* DH5α was used as host strain for high frequency transformation and plasmid DNA amplification (Sambrook, et al., 1989). Tryptophan, histidine and mevalonic acid lactone were dissolve in the same phosphate buffer as used in the culture medium.

Transformation procedure
Deletion plasmids were transformed to the protoplasts of *P. chrysogenum* strain DS54555 (Alvarez, et al., 1987). The phleomycin resistance gene was used as selection marker for the deletion of *roqA*, *roqT* and *roqD* while acetamidase gene (*amdS*) was used as selection marker for the deletion of *roqM*, *roqO*, *roqN* and *roqR* using acetamide as the only nitrogen source for selection (Kolar, et al., 1988; Kovalchuk, et al., 2012).

Genomic DNA extraction, total RNA extraction and cDNA amplification
Genomic DNA (gDNA) was isolated after 96 hours of growth on SMP medium using the modified yeast gDNA isolation protocol (Harju, et al., 2004) in which the fungal mycelium is broken in a FastPrep FP120 system (Qbiogene). Isolated gDNA was measured using a NanoDrop ND-1000 and 5 µg was used for southern hybridization. Total RNA of the host strain was isolated after 168 hours of growth in SMP medium using Trizol (Invitrogen), with additional DNase treatment using the Turbo DNA-free kit (Ambion). Total RNA was measured with the NanoDrop ND-1000 and a concentration of 500 ng per cDNA reaction was used. cDNA was synthesized using the iScript cDNA synthesis kit (Bio-Rad) in a 10-µl end volume.

Southern blot confirmation
gDNA of the host and various deletion strains was isolated using the E.Z.N.A. Fungal DNA kit (Omega Bio-tek). Southern blotting was carried out by digesting gDNA (5 µg) with the indicated restriction enzymes. Digested DNA fragments were separated on a 0.8 % agarose gel and blotted onto a Zeta-Probe membrane (Biorad) described earlier (Nijland, et al., 2008), and hybridized with the indicated probes that were DIG labeled.

qPCR analysis
The primers used to analyze the expression of all the genes in the Roquefortine/Meleagrin biosynthetic gene cluster i.e. Pc21g15480 (*roqA*), Pc21g15420 (*roqT*), Pc21g15430 (*roqD*), Pc21g15440 (*roqT*), Pc21g15450 (*roqN*), Pc21g15460 (*roqO*) and Pc21g15470 (*roqR*) were designed around an intron to avoid amplification on gDNA (Table S3). For expression analyses, the γ-actin gene was used as a control for normalization. A negative reverse transcriptase (RT) control was used to determine the gDNA contamination in isolated total RNA. The expression levels were determined in triplicate with a MiniOpticon system (Bio-Rad) using the Bio-Rad CFX manager software, with in which the threshold cycle (CT) values were determined automatically by regression. The SensiMix SYBR mix (Bioline) was used as a master mix for qPCR with 0.4 µM primers. The following thermocycler conditions were used: 95°C for 10 min, followed by 40 cycles of 95°C for 15 s, 60°C for 30 s, and 72°C for 30 s. Subsequently, a melting curve was generated to determine the specificity of the qPCRs.

Microarray methods
Triplicate shake flask cultivations were performed with the *P. chrysogenum* strain DS17690 to prepare a proprietary DNA microarray, using the Affymetrix Custom GeneChip program (Affymetrix): After 90 hours of cultivation, samples from shake flask cultures were filtered within seconds and quenched in liquid nitrogen. A standard protocol using Trizol reagent (Invitrogen) and acid phenol-chloroform was used to extract the total RNA followed by cDNA synthesis and cRNA synthesis. Hybridized arrays were scanned and analyzed using the Affymetrix GeneChip Operating Software (GCOS, Affymetrix) as described earlier by van der Berg et al. (van den Berg, et al., 2008).

HPLC-MS validation and analysis - Sample preparation
All strains used for gene assignments were grown in quintuplicate according to the procedure described above. Samples for determination of growth curves were grown in eight replicates whereas feeding experiment samples were grown in quadruplicate. Up to 50 µL of a thawed fermentation broth, 8 µL internal standard mixture containing 855 nmol/mL ranitidine, 657 nmol/mL reserpine and 1144 nmol/mL ampicillin was added. Subsequently, 230 µL of methanol was added for protein precipitation and vortexed for 10 minutes. The sample was then centrifuged at 14,000 g for 10 minutes at 10°C. 100 µL supernatant was transferred to an Eppendorf vial and evaporated for 30 minutes in a Thermo-Speedvac (Thermo Scientific, San Jose, CA). The dried sample was redissolved in 100 µL water containing 2 % acetonitrile, vortexed for 10 minutes and transferred to an autosampler vial.

Reversed-Phase LC-MS
For separation, an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled to a Surveyor PDA detector (Thermo Scientific, San Jose, CA) and LTQ-FT Ultra mass spectrometer (Thermo Scientific, San Jose, CA) were used. A sample of 5 µL was injected onto a Waters Atlantis T3 column (2.1 x 100 mm, 3 µm) (Waters, Milford, MA). The elution was performed with a linear gradient starting with 98 % of solvent A (1 % acetonitrile and 0.1 % formic acid in water) and 2 % solvent B (1 % water and 0.1 % formic acid in acetonitrile) for 1.5 minutes at a flow rate of 300 µL/min. The first linear gradient reached 40 % B at 22 minutes, the second 100 % B at 25 minutes. The column was flushed for 10 minutes at 100% B followed by equilibration for 8 minutes at 100 % A. The column effluent was directed to the ESI-LTQ-FT Ultra MS, operated in full scan (*m/z* 100-2000) in pos/neg switching mode with following settings: Positive ion mode (4kV source voltage, 14V capillary voltage, 65V tube lens), negative ion mode (3kV source voltage, -18V capillary voltage, -85V tube lens) with capillary temperature 275°C, sheath gas flow 50 and auxiliary gas flow 2.

Data Processing
Raw files were sliced into UV-trace, positive and negative mass trace and subsequently converted into NetCDF, using an in-house tool programmed in MATLAB (MathWorks, Natick, MA). NetCDF files with same polarity were subsequently deconvoluted in DataAnalysis 4.0 (Bruker Daltonik, Bremen, Germany) using the dissect function which was controlled through a macro. The resulting peak tables for each sample were combined and repeating features removed. The combined feature table was used as target list in which each feature was integrated in every individual sample. Samples were grouped regarding their biological origin and statistical tests were performed for determination of significant differences. Discovered features were selected and transferred to LCquan v2.6 (Thermo Scientific, San Jose, CA) for more accurate integration. Peaks were auto-integrated using base peak trace in a mass range of 10 ppm and retention time window of 60 seconds and manually corrected if necessary.

Secondary metabolite identification
The identity of **7** and **4** was confirmed by comparing retention time and HR-MS^n spectra of samples to commercially available standards. The structure of **1** was determined by Mass Spectrometry (MS^2 fragmentation and fragmentation tree comparison) as will be detailed elsewhere which was adapted from the metabolite identification pipeline developed with the Netherlands Metabolomics Centre (Kasper, et al., 2012; Rojas-Cherto, et al., 2012).
Compound **2**, **3** and **6** were identified by NMR. **3** and **6** were extracted based on a modified method of Ohmomo (Ohmomo, et al., 1980). A *P. chrysogenum* culture filtrate was made alkaline with 25 % ammonium hydroxide (pH 10) and extracted with dichloromethane. The alkaline dichloromethane layer was evaporated to dryness, redissolved in water containing 50 % acetonitrile, vortexed, centrifuged and transferred to an autosampler vial for fraction collection via preparative reversed phase LC on an Atlantis T3 column (10 x 100mm, 5 µm) (Waters Milford, MA). Compound **2** was extracted following the isolation procedure above except using ethylacetate as extraction solvent instead of dichloromethane.

NMR spectra were recorded on a Bruker Avance III 700 MHz NMR spectrometer, equipped with a 5 mm TCI probe. Samples were dissolved in 0.6 mL DMSO/$CDCl_3$ 50/50 and acquired at 280K (**2** and **3**) and 320K (**6**).

Results

Identification of roquefortine biosynthetic gene cluster

In order to identify the secondary metabolite biosynthetic genes under shake flask culture conditions, DNA microarray analysis was performed on the high penicillin yielding *P. chrysogenum* strain DS17690 grown in the presence and absence of the precursor phenylacetic acid (PAA). Most of the secondary metabolite gene clusters were not expressed or only at low levels (van den Berg, et al., 2008). However, high levels of expression were observed in the absence of PAA for a gene cluster including the nonribosomal peptide synthetase PC21g15480 (roqA), and associated genes Pc21g15420 (*roqT*), Pc21g15430 (*roqD*), Pc21g15440 (*roqN*), Pc21g15450 (*roqO*), Pc21g15460 (*roqM*) and Pc21g15470 (*roqR*) (Figure 1A). These genes were all down regulated when the cells were grown in the presence of PAA and hence are considered as co-regulated in the genome (Figure 1B). For the remainder of the study these genes were abbreviated as *roq* genes because of their involvement in roquefortine production.

Untargeted secondary metabolite profiling by HPLC-UV-MS

A robust and sensitive quantitative platform for profiling of secondary fungal metabolites from culture broth was developed, validated and applied. As the majority of these compounds are unknown and often present at low concentrations only, the method should be sensitive and provide a high degree of versatility. Herein, we used a high resolution HPLC-UV-MS method with positive/negative ionization switching in combination with UV to allow for the detection of several thousand features. To extract these features from the acquired data, a combination of commercial software packages and in-house scripts were used for untargeted peak discovery and automatic peak integration. By applying a deconvolution algorithm, fragments, adducts and cluster ions as well as their isotopes were grouped and altogether represented mainly by the most abundant ion in the resulting target list which facilitated further data analysis.

The method was validated in which several analytical performance characteristics were determined, but here merely the most important outcomes are reported. As secondary metabolites from the roquefortine/meleagrin pathway showed much higher ionization efficiency in the positive ion mode, only this polarity is described here. For the determination of linear dynamic range, sensitivity and reproducibility, retention times and signal intensities were evaluated for the used (internal) standards at multiple concentrations. These standards comprised of commercially available compounds from the meleagrin/neoxaline pathway (neoxaline, meleagrin and roquefortine C) and a non-related non-endogenous compound (ranitidine), which was used as internal standard. Retention time variations for the standards and several “unknown” endogenous compounds, which are spread over the entire chromatogram of 35 minutes, were limited to maximal 7 seconds (500 injections over 2 weeks of time). The method proved to be linear for each of the standards

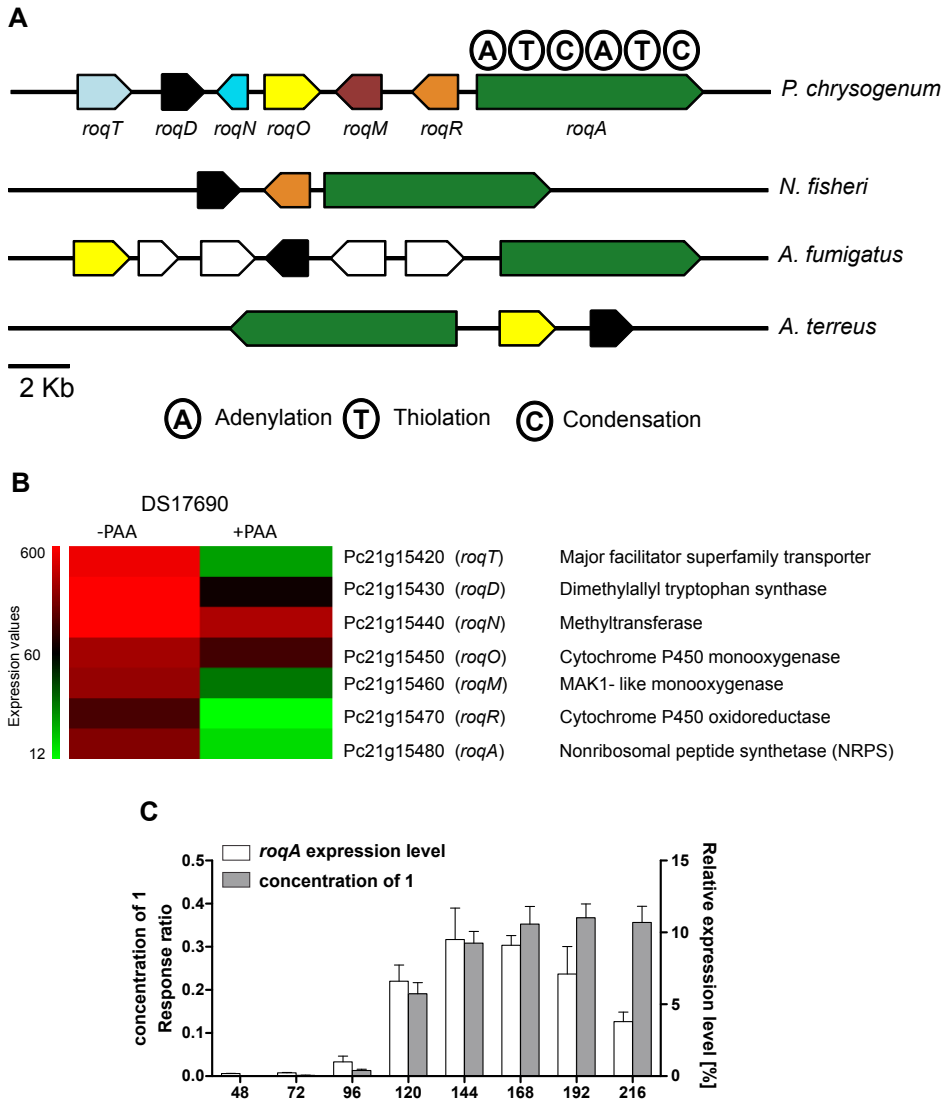


Figure 1. Organization of the roquefortine/meleagrin biosynthetic gene cluster and transcriptomic analysis.
A: Roquefortine/meleagrin biosynthetic gene cluster and their orthologs in phylogenetically relative species. Homologous proteins are indicated with the same color.
B: Microarray analysis of the roquefortine biosynthetic genes in *P. chrysogenum* DS54555 using shake flask culture conditions in the absence (-) or presence (+) phenylacetic acid (PAA).
C: Correlation between the expression level of *roqA* and the concentration of the product HTD (**1**) present in the growth media. The concentration of **1** was determined by HPLC-UV-MS.

over at least 3 orders of magnitude (R^2 ranged from 0.998 – 0.999 in both, academic and spiked matrix samples) in the appropriate concentration range. Associated Limit of Detections (LOD's) were determined from internal standard corrected calibration lines (8 levels) and ranged from 3-248 nM, depending on the compound of interest. Due to the absence of endogenous material in the roqA deletion strain used for validation, absolute quantitation at the reported low levels is realistic.
The recovery of the method was determined primarily by the extraction of the spiked compounds from the matrix, i.e., fermentation broth. Recoveries ranged from 88 % (rsd 7.8 %) to 55 % (rsd 2.9 %) depending on the compound. However, reproducibility measurements showed that both within day and between day reproducibility are well below 15 %. In summary, the analytical profiling platform developed here is characterized by good reproducibility and linearity, high coverage and high sensitivity reaching nanomolar levels.

Metabolite profiling of the culture broth of a P. chrysogenum strain with a deletion of the nonribosomal peptide synthetase gene roqA
In order to identify secondary metabolites synthesized by RoqA, its gene was deleted and comparative metabolite profiling was performed on the culture supernatant of the host and deletion strain. RoqA is specified by a 7.45 kbp long gene that translates into synthetase (2372 amino acids) with the typical domain motifs of nonribosomal peptide synthetases (NRPS). RoqA comprises two adenylation (A), thiolation (T) and condensation (C) domains arranged as ATCATC (Figure 1A) (van den Berg, et al., 2008). A host strain *P. chrysogenum* DS54555 was used which derived from a DS17690 strain lacking the *Ku70* gene and thus competent for homologous recombination. The DS54555 strain lacks all penicillin biosynthetic genes clusters, which facilitates the identification of unknown secondary metabolites in the culture broth as the profile is no longer dominated by β-lactam compounds. *RoqA* was deleted by homologous recombination using the deletion plasmid pDEST R4-R3a (Figure S1A) containing the phleomycin resistance gene surrounded by flanking regions of *roqA*. The deletion of *roqA* was confirmed by southern blot hybridization (Figure 2A). The host and *ΔroqA* strain were grown for 168 hours in secondary metabolite production medium (SMP Medium) and comparative metabolite analysis of these strains was carried out by HPLC-UV-MS (Figure S2). This revealed the loss of several secondary metabolites in the deletion strain as compared to the host strain. These metabolites were identified as histidyltryptophanyldiketopiperazine (HTD) (**1**), dehydrohistidyltryptophanyldiketo-piperazine (DHTD) (**2**), roquefortine D (**3**), roquefortine C (**4**), glandicoline B (**6**) and meleagrin (**7**). The structures of all compounds, except **1** were confirmed by LC-MS^2 analysis and subsequent fragmentation spectra, by retention time comparison with their extracted standards as well as by NMR (Figure 3, Figures S3, S4, S5 and S6, Table S1). The *m/z* value for the compound **1** identified as HTD was observed at 324.144 dalton representing the protonated molecule MH^+ with formula $C_{17}H_{18}N_5O_2$. Its chemical structure was elucidated using LC-MS^2 and high-resolution multi-dimensional fragmentation tree comparison. In addition, a protonated molecule MH^+ at *m/z* 404.171 and formula $C_{22}H_{22}N_5O_3$, which coincides with the protonated form of glandicoline A (**5**), was observed to be present in the host strain but absent in the deletion strain. Although, a full structure elucidation could not be performed, parts of the structure were identified by LC-

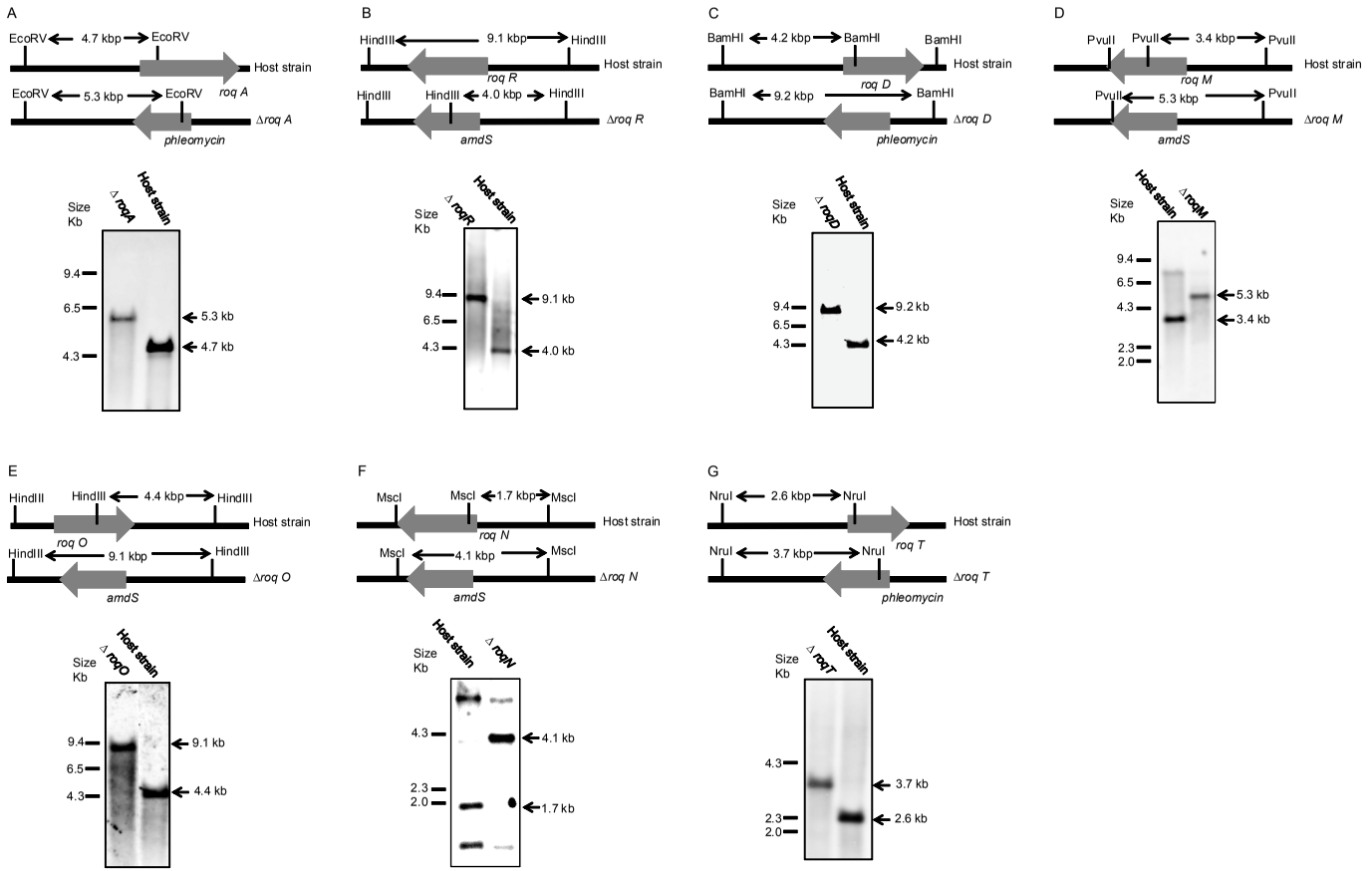


Figure 2. Southern blot analysis for deletion of the genes in the roquefortine/meleagrin pathway

Southern blot hybridization was performed with total DNA extracted from *P. chrysogenum* DS54555 strains with a deletion of the following genes: *roqA* (A), *roqR* (B), *roqD* (C), *roqM* (D), *roqO* (E), *roqN* (F) and *roqT* (G). The DNA was digested with the restriction enzymes as indicated in the schemes.

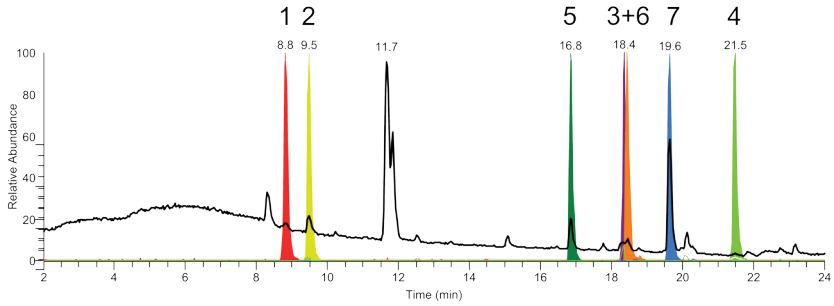


Figure 3. Total ion chromatogram for culture broth of *P. chrysogenum*
Total ion chromatogram (TIC, black) and normalized extracted ion chromatograms (EIC, colored) of secondary metabolites from the roquefortine/ meleagrin pathway in the culture broth of *P. chrysogenum* DS54555. HTD (**1**, 8.8 min), DHTD (**2**, 9.4 min), glandicoline A (**5**, 16.8 min), roquefortine D (**3**, 18.3 min), glandicoline B (**6**, 18.4 min), meleagrin (**7**, 19.6 min), roquefortine C (**4**, 21.4 min)

MS^2 which indicated consistency with the chemical structure of glandicoline A (e.g. dimethylallyl group) (Figure S3E). Hence, this ion was assumed to correspond to protonated glandicoline A. In conclusion, the loss of the identified secondary metabolites in the deletion strain proves unequivocally that RoqA is responsible for the initial reaction in the roquefortine/meleagrin pathway by combining L-tryptophan and L-histidine to 1.

Biochemical analysis of P. chrysogenum harboring a deletion of the roqR gene
In order to elaborate the putative associated genes of the roquefortine/meleagrin biosynthetic pathway, the *roqR* gene encoding a putative cytochrome P450 oxidoreductase was deleted. Herein, a linearized deletion plasmid pDEST R4-R3r (Figure S1B) was transformed into the protoplast of the host strain using the *Aspergillus nidulans* acetamidase gene (*amdS*) for the positive selection of transformants growing on media with acetamide as a sole nitrogen source. The homologous recombination resulted in the complete deletion of the *roqR* gene as demonstrated by Southern blot hybridization (Figure 2B). Both the host and *ΔroqR* strains were grown as described above. Deletion of *roqR* resulted in an accumulation of compound **1** and **3** while **2**, **4**, **5**, **6** and **7** were completely absent in the HPLC-MS profile of the *ΔroqR* strain (Figure 4A). These data suggest that **1** is a precursor of **2** resulting in a branch of the initially proposed linear pathway.

Biochemical analysis of P. chrysogenum with a deletion of the roqD gene
RoqD shares high sequence homology with dimethylallyltryptophan synthetase (DMATs). Transcriptional analysis of *roqD* showed high expression in the host strain grown in the absence of PAA in shake flasks (Figure 1B). To examine the role of RoqD in the biosynthesis of roquefortine and related metabolites, its gene was deleted with plasmid pDEST R4-R3d (Figure S1C) using phleomycin as a selection marker (Figure 2C). The deletion strain accumulated high levels of **1** (Figure 4B) in the culture broth as it was unable to add the dimethylallyl group needed for the conversion from **1** to **3**. Moreover, the biosynthesis of the other (downstream) metabolites **3**

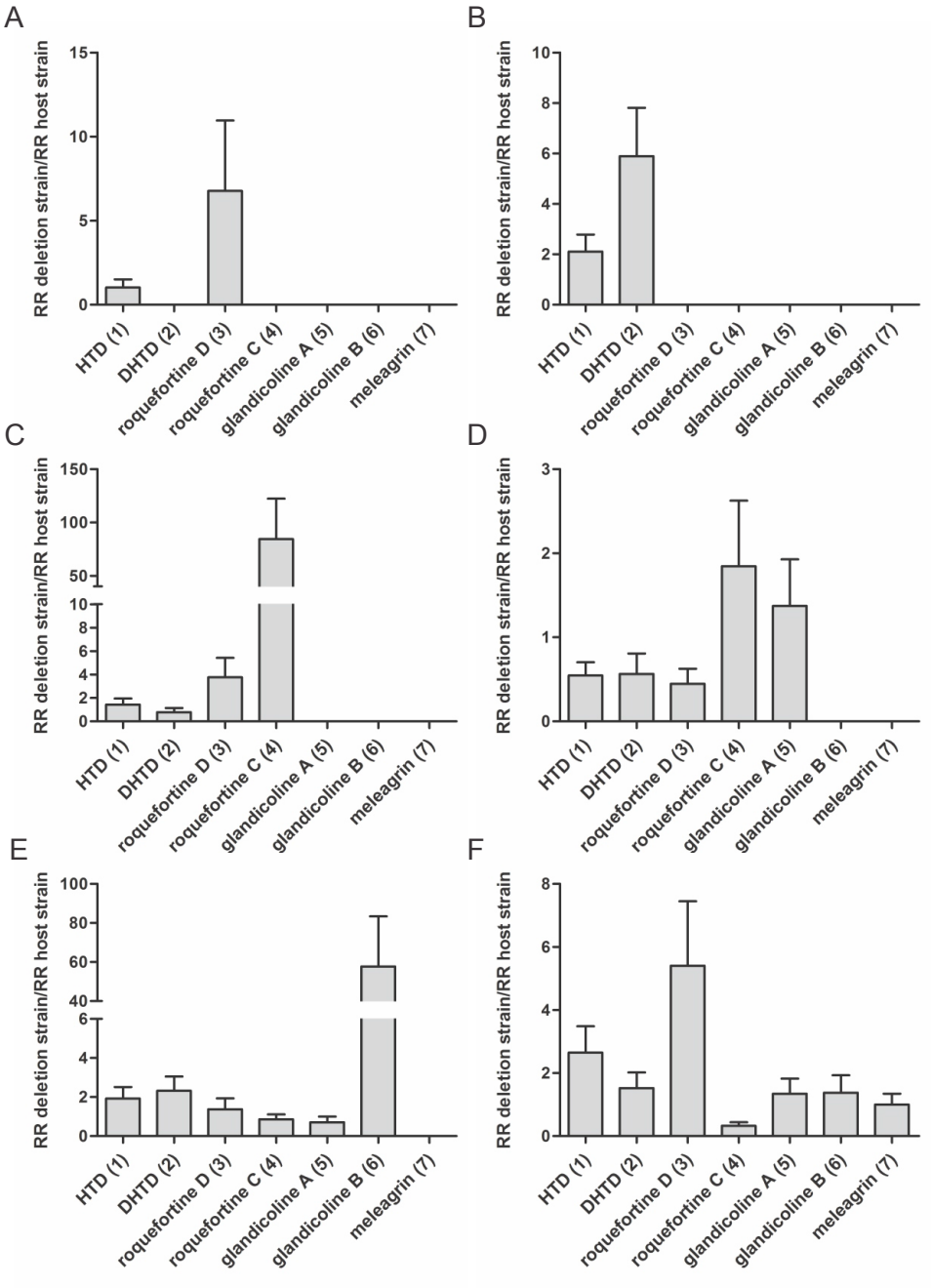


Figure 4. Internal standard corrected concentrations (RR = response ratio) of secondary metabolites from roquefortine/meleagrin pathway.
The metabolite concentrations in the culture broth of the *ΔroqR* (A), *ΔroqD* (B), *ΔroqM* (C), *ΔroqO* (D), *ΔroqN* (E) and *ΔroqT* (F) strains was compared to the host strain *P. chrysogenum* DS54555.

until **7** did not occur, while **2**, which is similar in structure to compound **1**, was six times higher in concentration as compared to the host strain. In combination with the *roqR* gene deletion, we conclude that RoqD is responsible for the conversion of **1** to **3** (see Discussion).

Biochemical analysis of P. chrysogenum strains with a deletion of roqM, roqN and roqO gene

RoqM shows homology to MAK1-like monooxygenases. Its gene was deleted with the deletion plasmid pDEST R4-R3m (Figure S1D) using *amdS* as selection marker (Figure 2D). The deletion of *roqM* led to substantial accumulation of **3** and **4** compared to the host strain whereas (downstream) metabolites like **5** till **7** were not detected (Figure 4C). This demonstrates that *roqM* is involved in the synthesis of **5** using **4**.

RoqO shows homology to cytochrome P450 monooxygenases, and its gene was deleted using plasmid pDEST R4-R3o (Figure 2E, Figure S1E) using *amdS* as selection marker. Metabolite profile comparison of the host and deletion strain showed similar concentrations of **5** and its upstream metabolites in both strains (Figure 4D). The absence of **6** and **7** in the deletion strain suggests **5** as final product and therefore indicates a role of *roqO* in the biosynthesis of **5** to **6**.

RoqN specifies a putative methyltransferase. pDEST R4-R3n (Figure S1F) was used for the deletion of *roqN* using *amdS* for the positive selection of transformants (Figure 2F). The deletion of *roqN* resulted in the loss of **7** in fermentation broth (Figure 4E). In addition, the concentration of **6** was increased sixty times compared to the host strain. This increase confirms that RoqN is responsible for the conversion of **6** to **7** (Garcia-Estrada, et al., 2011).

Biochemical analysis of P. chrysogenum harboring a deletion of the roqT gene

RoqT shows high sequence homology with transporters of the major facilitator superfamily. The gene was removed from the genome (Figure 2G) using deletion plasmid pDEST R4-R3t (Figure S1G) and phleomycin as selection marker. Remarkably, the strain with a deletion of *roqT* did not show any marked changes in the metabolite profile as compared to host strain (Figure 4F), except that the production of **3** was increased five times whereas production of **4** was decreased by 60 %. This indicates that *roqT* is not essential for the biosynthesis of **1** till **7**.

Gene expression and secondary metabolite production

To relate the extracellular metabolites of the roquefortine/meleagrin pathway to the expression levels of the biosynthetic genes, the host strain was grown for 216 hours in secondary SMP medium. Total mRNA extraction and extracellular metabolites analysis were carried with samples collected during growth. The metabolite concentrations were determined by HPLC-UV-MS, while transcript levels were determined by quantitative PCR using γ-actin as reference gene (Figure 1C, Figures S7and S8). The expression levels of the various biosynthetic genes increased linear in time, except for *roqA* that was found to be highly up-regulated after 96 hours when the cells are in the late log phase. The concentration of metabolite **1** synthesized by RoqA increased almost equally with the expression levels of *roqA* (Figure 1C), while after 192 hours the concentrations of metabolites **5** till **7** in the media

decreased (Figure S7 and S8). The concentration of **4** dropped significantly already after 168 hours. The production of all metabolites was particularly high after 96 hours of growth. It is concluded that the production of these metabolites is a late event during growth and that it correlates with the expression of the respective biosynthetic genes.

Precursor feed stimulates metabolite production

To evaluate the role of the predicted precursors in roquefortine/meleagrin pathway and related metabolites production, feeding experiments were performed. L-Tryptophan, L-histidine and mevalonic acid lactone were added individually and in combination (at high and low concentration, i.e., 30 and 10 mM respectively) to the cultures. Since the production of **7** and most of its derivatives is most significant around 96 hours of growth, precursors were added at that time point. Filtered supernatants were analyzed on HPLC-UV-MS after 168 hours of growth. The metabolites concentrations were dry weight corrected and compared via t-test to samples without precursors addition. The concentration of metabolites **1** till **7**, except **4** increased with the addition of 10 mM tryptophan (Figure 5). Addition of 30 mM tryptophan increased the production of **1**, **2**, **3**, **5** and **7** while addition of histidine, mevalonic acid lactone and combinations of tryptophan and histidine did not reveal any significant increase. These data suggest that the roquefortine biosynthetic pathway is limited by the availability of tryptophan.

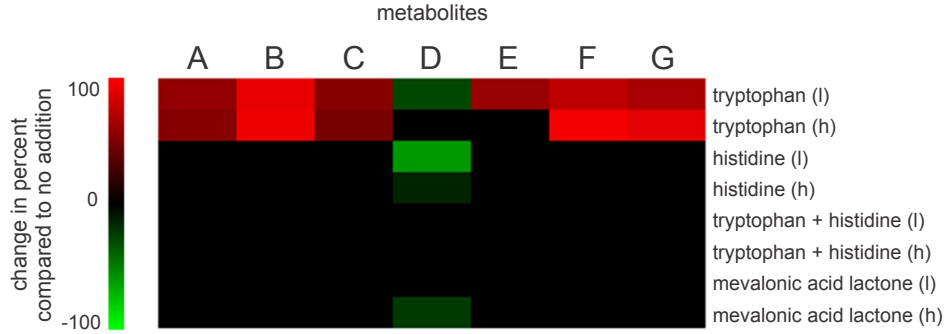


Figure 5. Change in production of roquefortine/meleagrin metabolites after precursor addition compared to production in cultures without addition. Colored cells show mean levels that are significant ($P < 0.05$) different.

Discussion

Here we have elucidated the biosynthetic pathway of *P. chrysogenum* responsible for the production of roquefortine, meleagrin and related compounds. Through a combined metabolic profiling, MS- and NMR based structure elucidation and gene inactivation analysis (7 genes in total, including a transporter-encoding gene); individual genes could be assigned to the various biosynthetic steps. The architecturally complex roquefortine and meleagrin scaffolds are synthesized from simple building blocks, i.e. histidine and tryptophan.

RoqA is a di-modulated NRPS, containing two adenylation domains (van den Berg, et al., 2008) responsible for the condensation of tryptophan and histidine to pro-

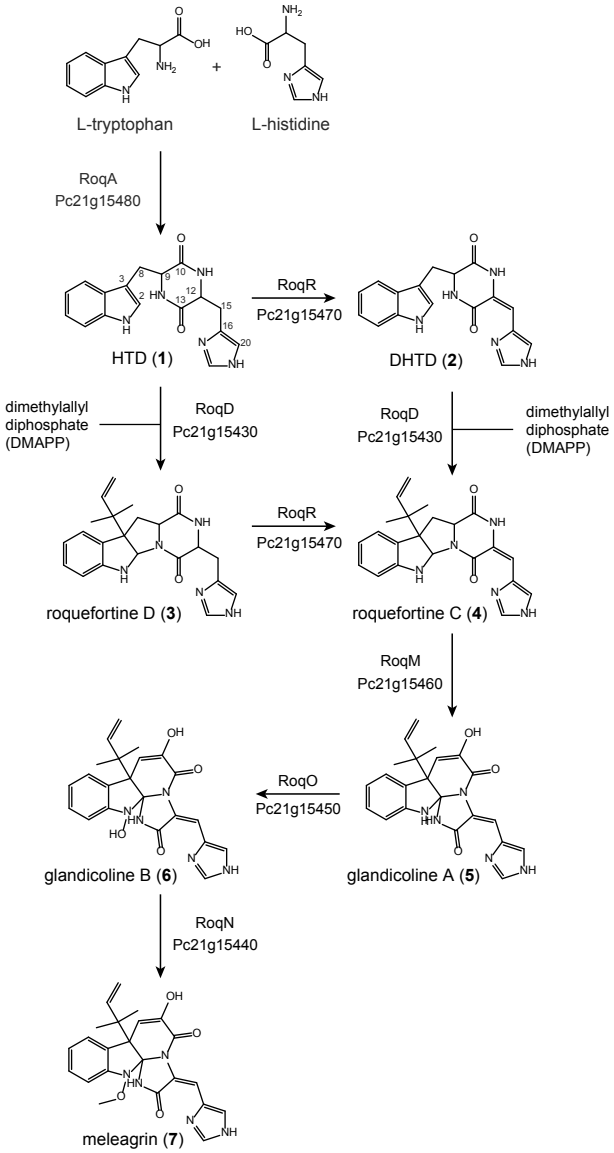


Figure 6. Proposed roquefortine/meleagrin biosynthetic pathway. See text for details.

duce **1**, a diketopiperazine. The *roqA* deletion strain no longer produced any of the roquefortine related metabolites (**1** till **7** absent from the broth) which was also recognized by García-Estrada et al. (Garcia-Estrada, et al., 2011). Therefore, *roqA* encodes the core enzyme initiating the biosynthesis of **1** till **7** and it is present in all species reported to produce these compounds (Frisvad, et al., 2004; Nielsen, et al., 2006). *RoqD* encodes a dimethylallyltryptophan synthase catalyzing the reversed prenylation of **1** at the C3 position in its indole moiety utilizing dimethylallyl diphosphate derived from mevalonic acid lactone of the mevalonate pathway to form **3** (Figure 6). In addition, it also catalyzes the ring closure between C2 and N14 of

the diketopiperazine moiety (Ohmomo, et al., 1975). Both conversions seem to occur simultaneously, similar to the formation of aszonalenin from benzodiazepinedione, as no intermediate could be observed (Yin, et al., 2009). No prenylated cyclohistidyltryptophanyl-diketopiperazine was detected in both the host and deletion strain as previously suggested (Garcia-Estrada, et al., 2011). The latter we attribute to a lack of convincing analytical evidence in that particular study that indeed this compound is formed. Importantly, this resulted in a assignment of RoqD and RoqM in the biosynthetic pathway. Oxidation of **1** at position C12 – C15 of the histidinyl moiety to **2** is carried out by the cytochrome p450 oxidoreductase encoded by *roqR* leading to a previously unknown branch in the pathway. In other studies, **2** was reported as a degradation product of **4** under acidic conditions and *in-vivo* (Scott, et al., 1979; Vinokurova, et al., 2001). However, the accumulation of **2** in the *ΔroqD* strain, despite the absence of **4** unequivocally indicates a different origin for the production of **2**. Two possible reactions of **1** lead to a branch of the roquefortine/meleagrin pathway, one via oxidation by RoqR to **2** and further to **4** by dimethylallyl addition by RoqD, and the other via dimethylallyl addition by RoqD to **3** and further to **4** via oxidation carried out by RoqR (Figure 6). **2** is therefore not just a degradation product of **4** but a true biological product of the roquefortine/meleagrin pathway. The absence of **3** and **4** in the *roqD* deletion strain and the absence of **2** and **4** in the *roqR* deletion strain strongly support our proposal that **2** is the precursor of **4** following an isoprene addition by RoqD at **2** to form **4**, similar to the known reaction of **1** to **3** (Figure 6). These observations are very intriguing as they question the precursor role of **3** to **4** in the proposed roquefortine/meleagrin pathway (Ohmomo, et al., 1975). **3** could be the end of the branch leaving **2** as single precursor for **4** if RoqD is unable of isoprene addition on **2** to yield **3**. Therefore, **2** or **3** or even both can be possible precursors of **4**.

Based on the highly significant accumulation of **4** in the *roqM* deletion strain and the absence of all downstream metabolites **5** till **7**, we concluded that *roqM*, encoding a MAK 1-monooxygenase like protein, is involved in the conversion of **4** to **5**. This leads to the conclusion that only one enzyme is involved in this prominent conversion which combines an oxidation, ring opening and ring closure in one reaction cascade (Steyn and Vleggaar, 1983). A possible mechanism involves an initial hydroxylation at C9 with a subsequent cleavage of the bond between C9 and N14, yielding a nine-membered ring with a keto-group at C9. Further oxidation at N1 of the indole moiety could lead to the loss of water to generate an imine which is followed by an attack of N11, ultimately yielding glandicoline A. Maackiain monooxygenases are known to be responsible for the minor conversion of Maackiain to 1-Hydroxymackian by the addition of an oxygen, which differs from the more prominent reaction from **4** to **5** found here (Covert, et al., 1996). It should however be emphasized that the enzyme assignment was initially based on sequence alignment only. Further definition of the mechanism requires an enzymological characterization.

Our results also demonstrate an involvement of RoqO in the conversion of **5** to **6**. *roqO* encodes a p450 monooxygenase that oxidizes **5** on its tryptophan moiety, resulting in formation of **6**. Remarkably, García-Estrada et al. (Garcia-Estrada, et al., 2011) mixed up the reactions catalyzed by RoqO and RoqM most likely because the lower concentrations of **6** were not recognized, whereas **5** could not have been

detected at all with the used HPLC-UV approach. Therefore, these authors wrongly assigned the biosynthetic pathway leading to the production of **5** by RoqO and **6** by RoqM. RoqN shares 99 % identity with the UbiE/COQ5 family methyltransferase in *A. origami*. Its activity was confirmed by inactivation of the *roqN* gene resulting in a loss of **7** concomitantly with a 60 times increase of **6**. Thus **7** is the methylated form of **6**, consistent with an involvement of methyltransferase as earlier recognized (Garcia-Estrada, et al., 2011).

Finally, *roqT* encodes a highly expressed member of the Multifacilitator Super Family of transporters that might be involved in active transport of some of these metabolites. However, its deletion had little effect on the metabolic profile suggesting that passive transport, diffusion or another transporter might be involved in secretion. It is remarkable that all of the described intermediates are found in the extracellular broth which suggests that these compounds effectively diffuse from the cell. Alternatively, secretion might be mediated by a relative unspecific export system that has not been identified so far. RoqT might fulfill a role in the retention of these metabolites (see also below) but this needs to be investigated further in feeding experiments.

In order to explore the evolutionary context of this gene cluster among filamentous fungi the roquefortine/meleagrin gene clusters in various fungi were compared using the antiSMASH algorithm (Medema, et al., 2011). Sorting the gene clusters by an empirical similarity score based on the number of BlastP hits between the predicted proteins of the gene clusters, gene order conservation and the percentage identity of the Blast hits resulted in three hits of gene clusters carrying more than two homologs of genes from the roquefortine/meleagrin biosynthesis gene cluster. One gene cluster, encoded in the genome of *Neosartorya fisheria*, contained three genes closely similar to genes of the *P. chrysogenum* roquefortine/meleagrin biosynthetic pathway (59, 67 and 60 % identity with *roqA*, *roqD* and *roqO*, respectively) (Figure 1A, Table S2). This suggests that *N. fisheria* may be capable of synthesizing compounds similar to **1**, **2**, **3** and **4**, using similar biosynthetic mechanisms. *A. fumigatus* and *A. terreus* also have three genes in common with the *P. chrysogenum* roquefortine gene cluster (*roqA*, *roqD* and *roqO*). Metabolites produced by these organisms are likely similar to **1** and **3**. However, the final structures may be different as for instance *A. fumigatus* contains at least four additional genes (annotated as phytanoyl-CoA dioxygenase, cytochrome P450, O-methyltransferase and cytochrome P450 monooxygenase) that might be involved in further modifications of these secondary metabolites (Table S2).

A time scale study for the production of roquefortine/meleagrin biosynthesis pathway intermediates and products showed an accumulation of **7** in the fermentation broth over time. After 192 hours of growth the metabolite concentrations reached their maximum before it declined due to a possible uptake from the media or degradation. Uptake of intermediates back into mycelia was previously reported for roquefortine (Kulakovskaya, et al., 1997) and it was proposed that these compounds serve as exogenous nitrogen source for colonial expansion (Overy, et al., 2005). As tryptophan, histidine and mevalonic acid lactone are the building blocks of the roquefortine/meleagrin pathway their effect on the production of **1**-**7** was determined in feeding experiments. Although labeling experiments showed the uptake and incorporation of these precursors (Barrow, et al., 1979) only the presence

of tryptophan stimulated secondary metabolite formation by this pathway (Figure 5). This implies that production is limited by the availability of tryptophan providing a lead for the optimization of roquefortine/meleagrin pathway by metabolic engineering of the shikimic acid or anthranilate route for tryptophan biosynthesis (Poulsen, et al., 1993). On the other hand, histidine reversed the effect of tryptophan. This might relate to nitrogen-dependent regulation of gene expression as roquefortine and related compounds have been implicated as exogenous nitrogen sources for colonial expansion (Overy, et al., 2005), and histidine is an excellent nitrogen source for *P. chrysogenum*. Finally, we are currently analyzing additional intermediates for structural elucidation. It is most likely that some of the enzymes of this pathway lack a certain degree of substrate specificity, which may lead to further branching and thus a wider palette of roquefortine/meleagrin related compounds.

Conclusion

Genome sequencing and microarray data analysis revealed seven genes clustered in the genome of *Penicillium chrysogenum* which are highly up regulated in the absence of the penicillin G precursor phenylacetic acid. These genes are involved in the biosynthesis of roquefortine/meleagrin metabolites. Complete deletion of all seven genes and detailed biochemical analysis of respective mutants strains via HPLC, MS and NMR revealed that dehydrohistidyltryptophanyldiketopiperazine (DHTD), which was previously identified as degradation product, is biologically synthesized and a precursor to Roquefortine C. Two enzymes of this pathway catalyze more than one reaction i.e. RoqD converts histidyltryptophanyldiketopiperazine (HTD) to roquefortine D and DHTD to roquefortine C while RoqR converts HTD to DHTD and roquefortine D to roquefortine C. Thus meleagrin is synthesized via a branched pathway rather than a linear biosynthesis.

Acknowledgements

This work was supported by the Perspective Genbiotics program subsidized by STW, and by the NWO-ACTS ibos program and (co)financed by the Netherlands Metabolomics Centre (NMC) which is a part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. HA was supported by HEC and STW. MR was supported by STW. We would like to thank Drs. H. Menke, W. Heijne and H. Roubos from the DSM Biotechology Centre for making available the DNA microarray data.

References

Alvarez, E., Cantoral, J.M., Barredo, J.L., Diez, B., and Martin, J.F. (1987). Purification to homogeneity and characterization of acyl coenzyme A:6-aminopenicillanic acid acyltransferase of Penicillium chrysogenum. Antimicrobial Agents and Chemotherapy 31, 1675-1682.

Aninat, C., Hayashi, Y., Andre, F., and Delaforge, M. (2001). Molecular requirements for inhibition of cytochrome p450 activities by roquefortine. Chemical research in toxicology 14, 1259-1265.

Barrow, K.D., Colley, P.W., and Tribe, D.E. (1979). Biosynthesis of the neurotoxin alkaloid Roquefortine. 225-226.

Challis, G.L. (2008). Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology

(Reading, England) 154, 1555-1569.

Covert, S.F., Enkerli, J., Miao, V.P., and VanEtten, H.D. (1996). A gene for maackiain detoxification from a dispensable chromosome of Nectria haematococca. Molecular & general genetics : MGG 251, 397-406.

Demain, A.L., and Fang, A. (2000). The natural functions of secondary metabolites. Advances in Biochemical Engineering/Biotechnology 69, 1-39.

Donadio, S., Sosio, M., Stegmann, E., Weber, T., and Wohlleben, W. (2005). Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. Molecular genetics and genomics : MGG 274, 40-50.

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences of the United States of America 105, 4601-4608.

Fleming, A. (2001). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. Bulletin of the World Health Organization 79, 780-790.

Frisvad, J.C., Smedsgaard, J., Larsen, T.O., and Samson, R.A. (2004). Mycotoxins, drugs and other extrolites produced by species in Penicillium subgenusPenicillium. STUDIES IN MYCOLOGY, 201-241.

Garcia-Estrada, C., Ullan, R.V., Albillos, S.M., Fernandez-Bodega, M.A., Durek, P., von Dohren, H., and Martin, J.F. (2011). A single cluster of coregulated genes encodes the biosynthesis of the mycotoxins roquefortine C and meleagrin in Penicillium chrysogenum. Chemistry & biology 18, 1499-1512.

Harju, S., Fedosyuk, H., and Peterson, K.R. (2004). Rapid isolation of yeast genomic DNA: Bust n' Grab. BMC biotechnology 4, 8.

Kasper, P.T., Rojas-Cherto, M., Mistrik, R., Reijmers, T., Hankemeier, T., and Vreeken, R.J. (2012). Fragmentation trees for the structural characterisation of metabolites. Rapid communications in mass spectrometry : RCM 26, 2275-2286.

Keller, N.P., Turner, G., and Bennett, J.W. (2005). Fungal secondary metabolism - from biochemistry to genomics. Nature reviews.Microbiology 3, 937-947.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology : FG & B 47, 736-741.

Kolar, M., Punt, P.J., van den Hondel, C.A., and Schwab, H. (1988). Transformation of Penicillium chrysogenum using dominant selection markers and expression of an Escherichia coli lacZ fusion gene. Gene 62, 127-134.

Kopp-Holtwiesche, B., and Rehm, H.J. (1990). Antimicrobial action of roquefortine. Journal of environmental pathology, toxicology and oncology : official organ of the International Society for Environmental Toxicology and Cancer 10, 41-44.

Kopp, B., and Rehm, H.J. (1981). Studies on the inhibition of bacterial macromolecule synthesis by roquefortine, a mycotoxin from Penicillium roqueforti. European journal of Applied Microbiology and Biotechnology 13, 232-235.

Kovalchuk, A., Weber, S.S., Nijland, J.G., Bovenberg, R.A., and Driessen, A.J. (2012). Fungal ABC transporter deletion and localization analysis. Methods in molecular biology (Clifton, N.J.) 835, 1-16.

Kulakovskaya, T.V., Reshetilova, T.A., Kuvichkina, T.N., and Vinokurova, N.G. (1997). Roquefortine excretion and uptake by Penicillium crustosum Thom VKM F- 1746. Process Biochemistry 32, 29-33.

Ma, L.J., van der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464, 367-373.

Maristela, B.M., and Carvalho, I. (2007). Diketopiperazines: biological activity and synthesis. Tetrahedron, 9923-9932.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic acids research 39, W339-346.

Nielsen, K.F., Sumarah, M.W., Frisvad, J.C., and Miller, J.D. (2006). Production of metabolites from the Penicillium roqueforti complex. Journal of Agricultural and Food Chemistry 54, 3756-3763.

Nijland, J.G., Kovalchuk, A., van den Berg, M.A., Bovenberg, R.A., and Driessen, A.J. (2008). Expression of the transporter encoded by the cefT gene of Acremonium chrysogenum increases cephalosporin production in Penicillium chrysogenum. Fungal genetics and biology : FG & B 45, 1415-1421.

Ohmomo, S., Ohashi, T., and Abe, M. (1980). Isolation of biogenetically correlated 4 alkaloids from the cultures of Penicillium corymbiferum. Agric. Biol. Chem. 44, 1929-1930.

Ohmomo, S., Sato, T., Utagawa, T., and Abe, M. (1975). Isolation of festuclavine and three new indole alkaloids, roquefortine A, B and C from the cultures of Penicillium roqueforti. Agricultural and Biological Chemistry 39, 1333-1334.

Overy, D.P., Nielsen, K.F., and Smedsgaard, J. (2005). Roquefortine/oxaline biosynthesis pathway metabolites in Penicillium ser. Corymbifera: in planta production and implications for competitive fitness. J Chem Ecol 31, 2373-2390.

Polonsky, J., Merrien, M.A., and Scott, P.M. (1977). Roquefortine and isofumigaclavine A, alkaloids from Penicillium roqueforti. Annales de la Nutrition et de l'Alimentation 31, 963-968.

Poulsen, C., Bongaerts, R.J., and Verpoorte, R. (1993). Purification and characterization of anthranilate synthase from Catharanthus roseus. European journal of biochemistry / FEBS 212, 431-440.

Rojas-Cherto, M., Peironcely, J.E., Kasper, P.T., van der Hooft, J.J., de Vos, R.C., Vreeken, R., Hankemeier, T., and Reijmers, T. (2012). Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. Analytical Chemistry 84, 5524-5534.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press. second ed.

Scherlach, K., and Hertweck, C. (2006). Discovery of aspoquinolones A-D, prenylated quinoline-2-one alkaloids from Aspergillus nidulans, motivated by genome mining. Organic & biomolecular chemistry 4, 3517-3520.

Scott, P.M., J.Polonsky, and M.A.Merrien. (1979). Configuration of the 3,12 Double Bond of Roquefortine. J. Agric. Food Chem. 27, 201-202.

Scott, P.M., and Kennedy, P.C. (1976). Analysis of blue cheese for roquefortine and other alkaloids from Penicillium roqueforti. Journal of Agricultural and Food Chemistry 24, 865-868.

Shwab, E.K., and Keller, N.P. (2008). Regulation of secondary metabolite production in filamentous ascomycetes. Mycological Research 112, 225-230.

Steyn, P.S., and Vleggaar, R. (1983). Roquefortine, an intermediate in the biosynthesis of oxaline in cultures of Penicillium oxalicum. J Chem Soc Chem Commun. 10, 560-561.

van den Berg, M.A., Albang, R., Albermann, K., Badger, J.H., Daran, J.M., Driessen, A.J., Garcia-Estrada, C., Fedorova, N.D., Harris, D.M., Heijne, W.H., et al. (2008). Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. Nature biotechnology 26, 1161-1168.

Vinokurova, N.G., Zelenkova, N.F., Baskunov, B.P., and Arinbasarov, M.U. (2001). Determination of Diketopiperazine Alkaloids of the Roquefortine Group by UV Spectroscopy, Thin-Layer Chromatography and High-Performance Liquid Chromatography. Zh. Anal. Khim 56, 258-262.

Weber, S.S., Bovenberg, R.A., and Driessen, A.J. (2012). Biosynthetic concepts for the production of beta-lactam antibiotics in Penicillium chrysogenum. Biotechnology journal 7, 225-236.

Yin, W.B., Grundmann, A., Cheng, J., and Li, S.M. (2009). Acetylaszonalenin biosynthesis in Neosartorya fischeri. Identification of the biosynthetic gene cluster by genomic mining and functional proof of the genes by biochemical investigation. The Journal of biological chemistry 284, 100-109.

Supplemental Data

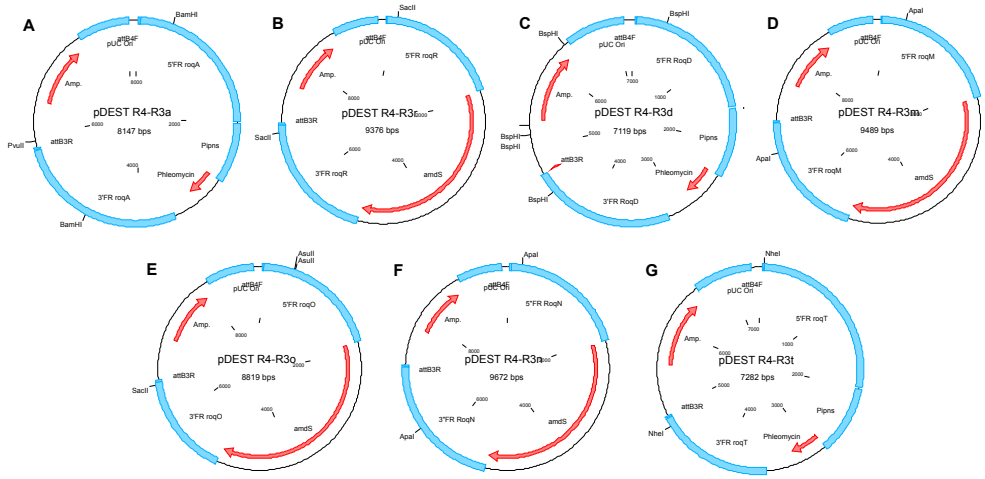


Figure S1. Map of the deletion constructs for *roqA*, *roqR*, *roqD*, *roqM*, *roqO*, *roqN* and *roqT* which were used for deletion. Features of the vectors: Amp, Ampicillin resistance gene for the selection in *E. coli* ; *ori*, pUC origin of replication; attP3, and attP4, Gateway *att* recombination sites; Pipns, promoter of *P. chrysogenum pcbC* gene; Phleomycin, resistance gene for selection in fungi. *amdS*, *A. nidulans* acetamidase gene.

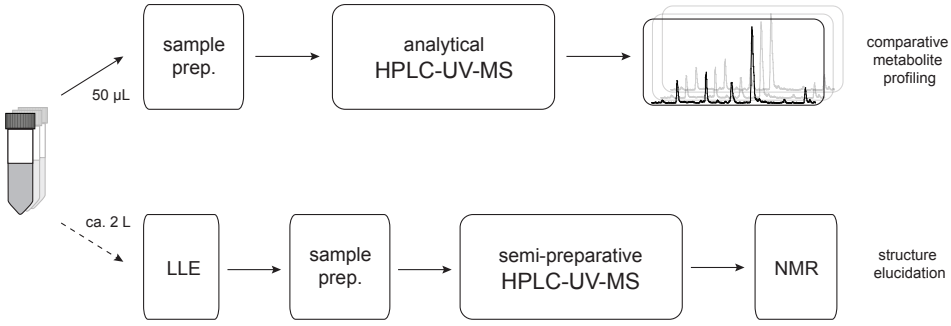


Figure S2. Analytical approach in schematic view. Proteins, present in fermentation broth, were removed during sample preparation. Samples were analyzed by HPLC-UV-MS and comparative metabolite profiling performed. Statistical significant features were extracted using liquid-liquid extraction (LLE) and semi-preparative HPLC-UV-MS and their structure elucidated by NMR.

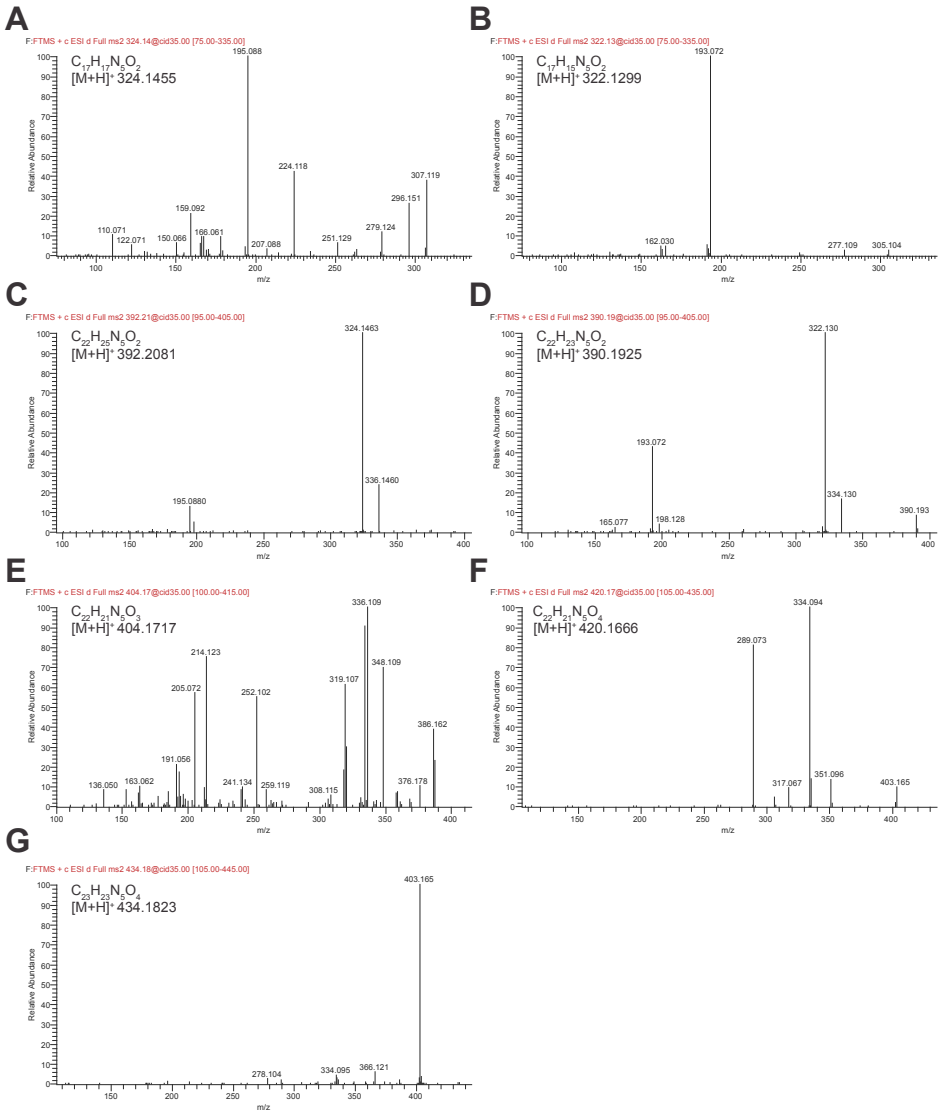


Figure S3. HPLC-MS^2 fragmentation spectra including chemical formula and calculated exact mass of the protonated HTD (A), DHTD (B), roquefortine D (C), roquefortine C (D), glandicoline A (E), glandicoline B (F) and meleagrin (G) acquired at LTQ-FT-MS Ultra at 35% normalized collision energy in positive mode.

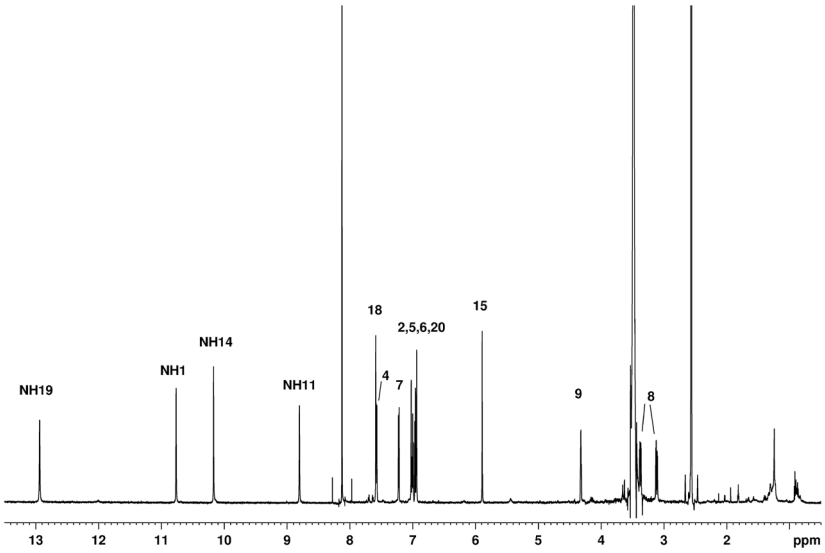


Figure S4. 1H NMR spectrum of DHTD (**2**).

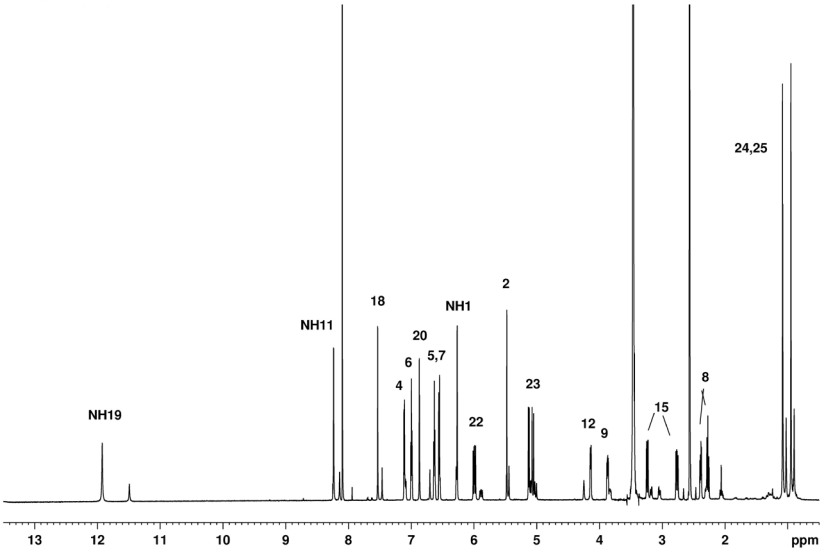


Figure S5. ^{1}H NMR spectrum of roquefortine D (**3**). Small additional peaks are not due to impurities but to a second conformation of roquefortine D.

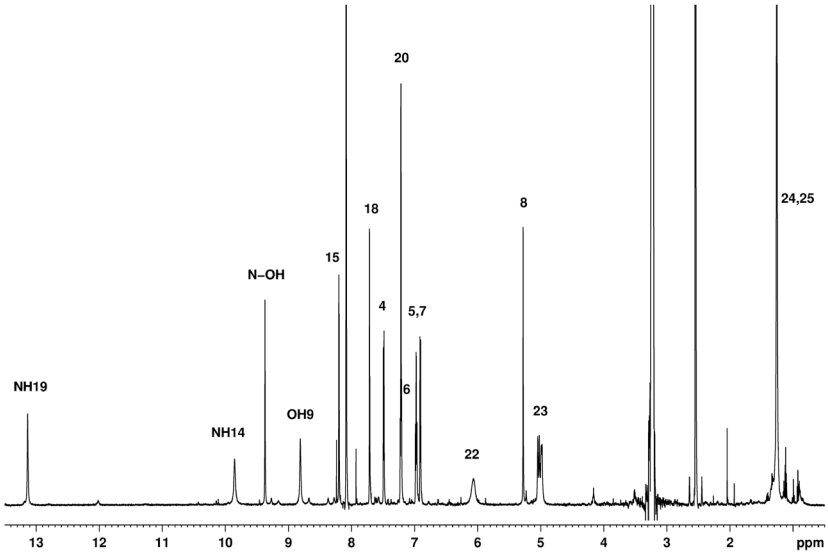


Figure S6. ^{1}H NMR spectrum of glandicoline B (**6**).

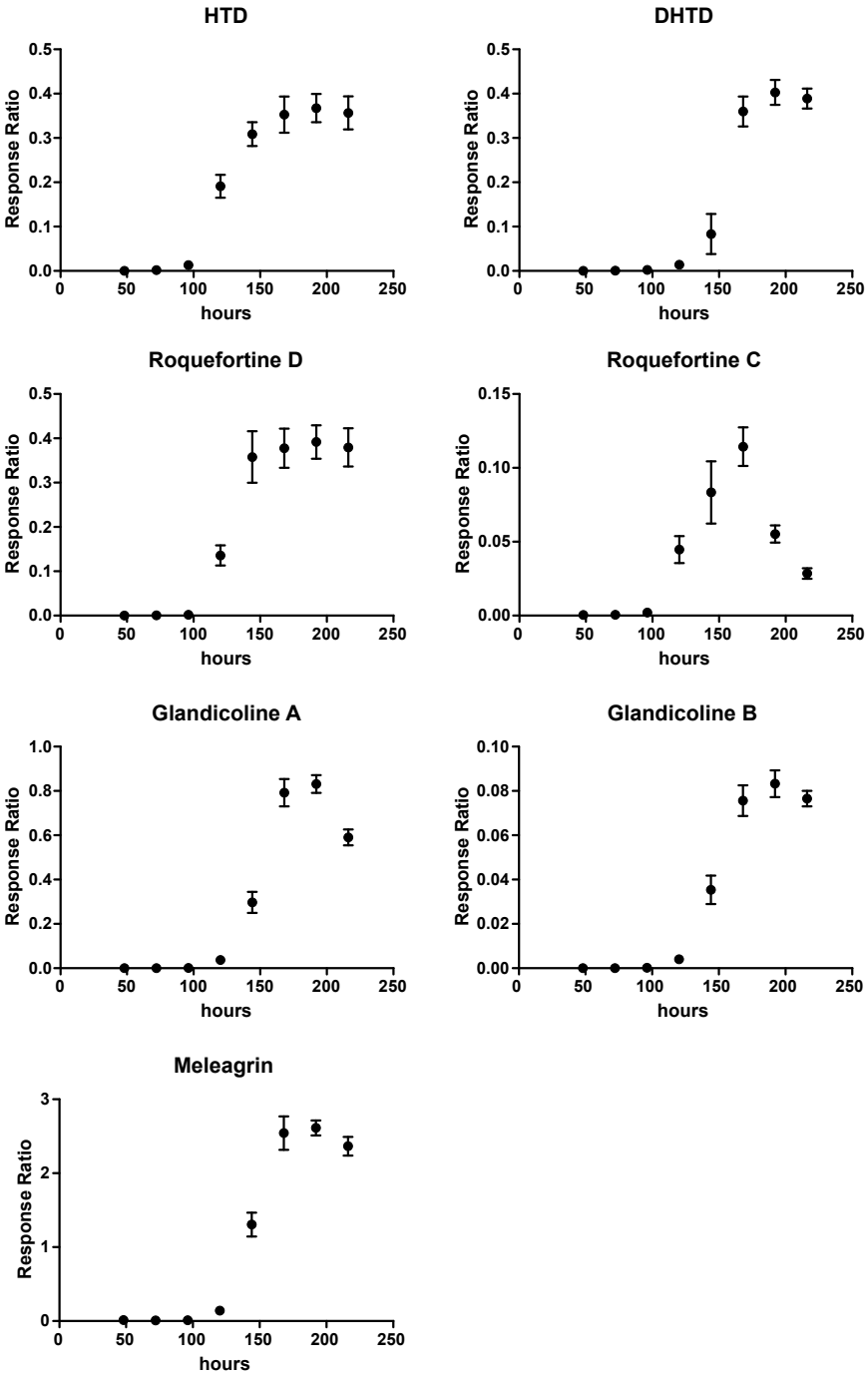


Figure S7. Internal standard corrected metabolite concentration in fermentation broth of *P. chrysogenum* AFF393 sampled at multiple time points and determined by HPLC-UV-MS.

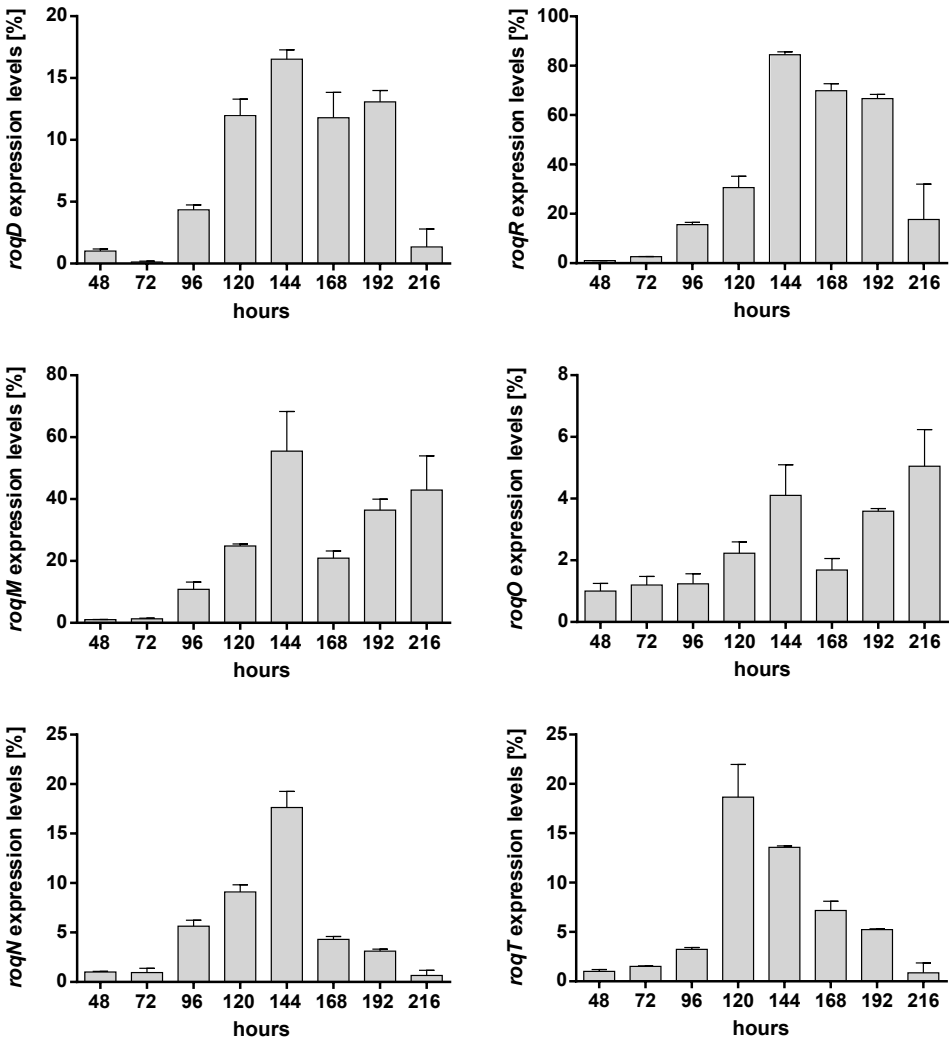


Figure S8. Temporal expression of roquefortine/meleagrin biosynthetic gene cluster in *P. chrysogenum* AFF393 grown in shaking flask culture.

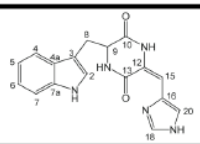
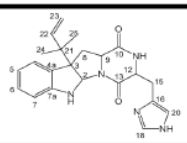
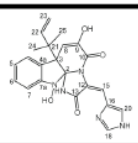
	DHTD (**2**) TE=280K			roquefortine D (**3**) TE=280K			glandicoline B (**6**) TE=320K	
								
	^{1}H (δ)	^{13}C (δ)		^{1}H (δ)	^{13}C (δ)		^{1}H (δ)	^{13}C (δ)
NH1	10.77	n.a.	NH1	6.27	n.a.	NOH1	9.40	n.a.
C2	7.03	124.4		5.48	76.7		n.a.	101.7
C3	n.a.	106.9		n.a.	60.7		n.a.	52.5
C4	7.58	118.6		7.12	124.5		7.51	124.1
C4a	n.a.	127.4		n.a.	128.6		n.a.	126.5
C5	6.96	118.4		6.64	117.4		6.99	122.0
C6	7.00	120.7		7.00	128.3		7.23	127.6
C7	7.22	111.0		6.56	108.5		6.93	111.4
C7a	n.a.	136.1		n.a.	150.8		n.a.	148.2
C8	3.38 3.12	29.6		2.39 2.28	36.5		5.29	109.0
C9	4.32	56.0		3.87	58.2		n.a.	142.8
OH9	n.a.	n.a.		n.a.	n.a.		8.86	n.a.
C10	n.a.	165.8		n.a.	168.2		n.a.	158.8
NH11	8.81	n.a.		8.24	n.a.		n.a.	n.a.
C12	n.a.	121.6		4.14	55.2		n.a.	124.4
C13	n.a.	160.9		n.a.	164.9		n.a.	166.5
NH14	10.17	n.a.		n.a.	n.a.		9.90	n.a.
C15	5.90	107.4		2.77 3.24	28.0		8.20	106.8
C16	n.a.	125.5		n.a.	136.3		n.a.	126.1
NH17	12.91	n.a.		11.93	n.a.		13.16	n.a.
C18	7.59	135.8		7.54	134.8		7.73	136.9
C20	6.94	133.0		6.87	113.3		7.23	133.5
C21	n.a.	n.a.		n.a.	40.6		n.a.	41.8
C22	n.a.	n.a.		6.00	143.7		6.08	143.3
C23	n.a.	n.a.		5.13 5.06	113.9		5.05 5.00	112.6
C24	n.a.	n.a.		1.09	22.1		1.28	23.4
C25	n.a.	n.a.		0.95	22.6		1.28	23.4

Table S1. ^{1}H and ^{13}C NMR chemical shifts of DHTD (**2**), roquefortine D (**3**) and glandicoline B (**6**) in DMSO/$CDCl_3$ at 320K and 280K, respectively (δ in ppm).

Neosartorya fischeri NRRL 181

Query gene	Subject gene	% Identity	Blast score	%Coverage	E-value
roq D	NFIA_074280	67	602	100	0.0
roq R	NFIA_074290	60	598	100	0.0
roq A	NFIA_074230	59	2644	92	0.0

Aspergillus fumigatus Af293

Query gene	Subject gene	% Identity	Blast score	%Coverage	E-value
roq D	AFUA_8G00210	31	206	100	1e-62
roq O	AFUA_8G00240	65	646	94	0.0
roq A	AFUA_8G00170	32	955	89	0.0

Aspergillus terreus NIH2624

Query gene	Subject gene	% Identity	Blast score	%Coverage	E-value
roq D	ATEG_10306	31	226	100	7e-71
roq O	ATEG_10307	47	333	70	5e-112
roq A	ATEG_10305	32	920	81	0.0

Table S2. BlastP analysis of roquefortine/meleagrine biosynthetic pathway genes.

Target	Primer sequence (5'- 3') Forward	Reverse
roqA	AATTAGTGGCTTCATCTCC	CGGGTGATATACTGCAGTCC
roqD	CTTGGTCGGCATTCCCGAGC	ATAGTACATGGTGAGGTATGG
roqR	CCTGCGCAATACACTGGCGG	TGACACGGCTCCTGAATCATGG
roqM	CTCGCATCTGACTATAAATCGC	CTCGCAGACTACAAGATCATC
roqO	GACGACGATTGCTGACACC	CATGGTTATGCAGCGAGCC
roqN	CAGTCCACTCCTGTGGCACC	GAATTCATGTCCTGTATGAACC
roqT	AACTGATCCTCTACCGCAGG	GTGAGTCGAACTGATCTGTG

Table S3. Primers designed for gene expression analysis of roquefortine/meleagrine biosynthetic gene.

Chapter 3

Novel key metabolites reveal further branching of the roquefortine/meleagrin biosynthetic pathway

Based on

Marco I. Ries*, Hazrat Ali*, Peter P. Lankhorst, Thomas Hankemeier, Roel A.L. Bovenberg, Arnold J.M. Driessen, Rob J. Vreeken
Novel key metabolites reveal further branching of the roquefortine/meleagrin biosynthetic pathway
Submitted for publication

** these authors contributed equally*

Abstract

Metabolic profiling and structural elucidation of novel secondary metabolites obtained from derived deletion strains of the filamentous fungus Penicillium chrysogenum were used to reassign various previously ascribed synthetase genes of the roquefortine/meleagrin pathway to their corresponding products. Next to the structural characterization of roquefortine F and neoxaline, which are for the first time reported for P. chrysogenum, we identified three novel metabolites, namely roquefortine L, M and N which harbor remarkable chemical structures. Their biosynthesis is discussed, questioning the exclusive role of glandicoline A as key intermediate in the pathway. The results reveal that further enzymes of this pathway are rather unspecific and catalyze more than one reaction leading to excessive branching in the pathway with meleagrin and neoxaline as end products of two branches.

Introduction

The filamentous fungus Penicillium chrysogenum is commercially exploited for many decades due to its high production of β-lactam antibiotics like penicillin G (Weber, et al., 2012). Next to penicillins, secondary metabolites like roquefortines and glandicolines were isolated from liquid cultures of P. chrysogenum which show pharmaceutically interesting properties, like neurotoxic (Scott, et al., 1976), antimicrobial (Clark, et al., 2005; Koolen, et al., 2012) and antitumor (Du, et al., 2010) activities. They are structurally closely related and arise from the roquefortine/meleagrin pathway which contains a di-modular Non-Ribosomal Peptide Synthetase (NRPS) flanked by six associated genes (Ali, et al., 2013; Garcia-Estrada, et al., 2011). Starting with histidyltryptophanyldiketopiperazine (HTD), synthesized by the core synthetase enzyme RoqA using tryptophan and histidine as substrates, RoqD catalyzes the reversed prenylation of HTD at the C-3 of its indole moiety utilizing dimethylallyldiphosphate to form roquefortine D. At the same time, RoqR, a cytochrome p450 oxidoreductase, oxidizes HTD at its histidinyl moiety to dehydrohistidyltryptophanyldiketo-piperazine (DHTD). Both simultaneous reactions of HTD lead to a branch of the roquefortine/meleagrin pathway, one to DHTD via the oxidation by RoqR and further to roquefortine C by dimethylallyl addition of RoqD, and the other via an alteration of the enzymatic order. There, dimethylallyl addition is first performed by RoqD to yield roquefortine D while further oxidation is carried out by RoqR yielding roquefortine C (Figure 1). Although several labeling, silencing and deletion experiments have been conducted, there is still ambiguity about the subsequent biosynthetic reactions and the genes involved. For instance, roquefortine C is supposed to be converted into glandicoline A and further to glandicoline B with RoqM and RoqO each catalyzing one reaction (Ali, et al., 2013; Garcia-Estrada, et al., 2011). However, their assignment to a particular reaction is still unclear. In addition, neoxaline was proposed as final product of the pathway, originating from a hydrogenation of meleagrin (Overy, et al., 2005), yet no gene could be found in the roq gene cluster performing that reaction.
Here we describe the quantification, structural identification and biosynthesis of five previously unidentified metabolites, obtained from high sensitive comparative metabolite profiling of host and deletion strains. Roquefortine F and neoxaline, next to the three structurally novel metabolites, which we named roquefortine L, M and N, were found to be derived from the roquefortine/meleagrin pathway. These results demonstrate a further branching of this secondary metabolite pathway yielding a variety of intermediates with complex structures and a diverse range of activities.

Experimental procedures

Host strains, media, grown condition and plasmid construction
P. chrysogenum strain DS54555, which lacks both the penicillin cluster genes and the ku70 gene, was used as a host strain for deletion analysis and was kindly supplied by DSM Anti-infective (Delft, the Netherlands). All the strains were grown on YGG-medium for protoplasts formation and transformation. For analysis, cells were grown on SMP medium (glucose, 5.0 g/L; lactose, 75 g/L; urea, 4.0 g/L; Na_2SO_4, 4.0 g/L; CH_3COONH_4, 5.0 g/L; K_2HPO_4, 2.12 g/L; KH_2PO_4, 5.1 g/L) for secondary metabo-

lites production using a shaking incubator at 200 rpm for 168 hours at 25°C.

Metabolite profiling
All strains used for gene assignments were grown in quintuplicate to increase statistical power, according to the procedure described above. Sample preparation was carried out as described previously (Ali, et al., 2013). Metabolomic profiling was performed on an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled to a Surveyor PDA detector (Thermo Scientific, San Jose, CA) and LTQ-FT-ICR-Ultra mass spectrometer (Thermo Scientific, San Jose, CA) equipped with an ElectroSpray Interface as described earlier (Ali, et al., 2013).

Metabolite identification
The identity of compound **10** was confirmed by comparing retention time and MS fragmentation spectra to its commercially available standard, purchased from Bio-Connect (Huissen, the Netherlands). Compound **6**, **9**, **11** and **12** were identified using NMR after extraction from liquid cultures. **6** was extracted from the *roqN* deletion strain culture filtrate which was made alkaline with 25 % ammonium hydroxide (pH 10) and extracted with dichloromethane. The alkaline dichloromethane layer was evaporated to dryness, redissolved in water containing 50 % acetonitrile, vortexed, centrifuged and transferred to an autosampler vial for fraction collection via preparative reversed phase LC on an Atlantis T3 column (10 x 100mm, 5 µm) (Waters Milford, MA). Compound **9** was extracted following the isolation procedure above except using culture filtrate of the *roqO* deletion strain, while **11** and **12** were obtained from the same culture filtrate after lyophilization and extraction using methanol. The methanol layer was evaporated to dryness, redissolved in water, vortexed, centrifuged and subjected to repeated semi-preparative chromatography as described above. Elemental composition of compounds **6**, **9**, **11** and **12** was determined using high-resolution MS. NMR spectra were recorded on a Bruker Avance III 700 MHz or 600 MHz spectrometer with sample temperatures ranging from 260 K to 300 K, depending on the particular requirements for each sample. By choosing an optimal aquisation temperature, severe line broading could be avoided which was observed for various signals due to conformational averaging. For aquisition, samples were dissolved in equal amounts of DMSO and $CDCl_3$.

*Chemical stability of compound **6***
An aqueous solution of compound **6** was adjusted to pH 2.5 by addition of formic acid. Metabolite profiling was carried out as described above. Products, formed by a degradation of **6**, were compared to extracted standards using HPLC-MS/MS.

Results

Metabolite profiling of host and deletion strains leads to five new metabolites of the roquefortine/meleagrin pathway
In a previous study, we described the identification of various abundant metabolites and resolved the major enzymatic steps belonging to the roquefortine/meleagrin pathway (Ali, et al., 2013). In order to identify secondary metabolites originating from the *roq* gene cluster (Figure 1A), culture supernatants of the host strain and

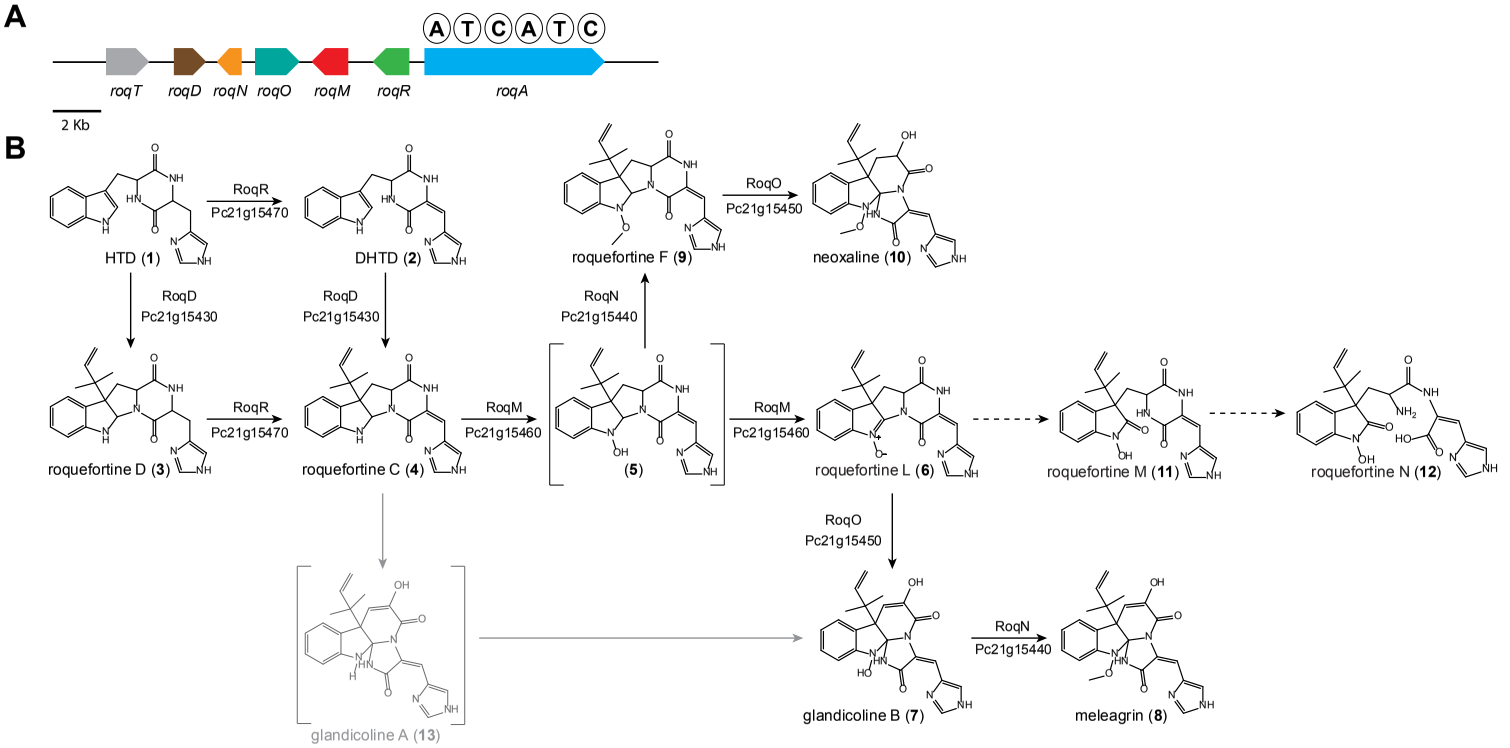


Figure 1. Roquefortine/meleagrin biosynthetic gene cluster and proposed corresponding pathway.
(A) Organization of the roquefortine/meleagrin biosynthetic gene cluster. (B) Proposed roquefortine/meleagrin pathway. Numbers between brackets are compound identifiers used throughout the manuscript. Enzymatic catalyzed reactions are indicated by solid arrows whereas chemical reactions are indicated by dashed arrows. Structures shown in brackets could not be detected whereas grey colored reactions and compounds were previously proposed for various Penicillium species (Garcia-Estrada, et al., 2011; Overy, et al., 2005; Vinokurova, et al., 2002).

individually *roq* gene deletion strains were subjected to comparative metabolite profiling using HPLC-UV-MS (Figure 2). As host strain, *P. chrysogenum* DS54555 was used which is derived from the industrial DS17690 strain lacking the *ku70* gene and multiple penicillin biosynthetic genes clusters. Here, we describe the identification and quantification of several less abundant metabolites, roquefortine L (**6**), roquefortine F (**9**), neoxaline (**10**), roquefortine M (**11**) and roquefortine N (**12**) (Figure 1B) that have not been previously considered or structurally characterized, filling missing biosynthetic reaction steps in the roquefortine/meleagrin pathway.

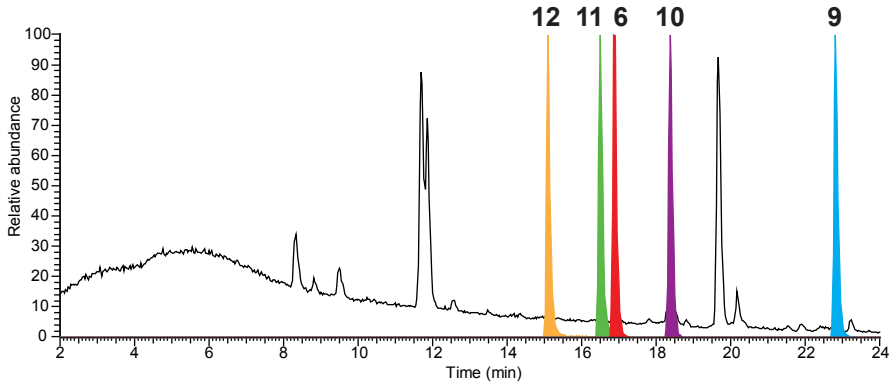


Figure 2. HPLC-MS elution profiles of novel metabolites of the meleagrin/neoxaline pathway.
HPLC-MS total ion chromatogram (TIC, black) and normalized extracted ion chromatograms (EIC, colored) of novel secondary metabolites from the meleagrin/neoxaline pathway. Roquefortine N (**12**, 15.1 min), roquefortine M (**11**, 16.5 min), roquefortine L (**6**, 16.8 min), neoxaline (**10**, 17.8 min), and roquefortine F (**9**, 22.8 min).

*Structure elucidation and quantification of **6**, **11** and **12***

Compound **6** is a novel complex metabolite composed of a roquefortine scaffold and a rare nitrone moiety, thus named roquefortine L. The mass-to-charge ratio of its corresponding ion was observed at 404.1706 using HPLC-FT-ICR-MS, representing the protonated molecule $[M+H]^+$ with formula $C_{22}H_{22}N_5O_3$ (calc. 404.1717) eluting at 16.5 minutes. The same ion was previously tentatively identified as glandicoline A (**13**) (Figure 1B), as elemental composition and parts of the structure indicated consistency with this compound (Ali, et al., 2013). However, its ^{1}H- and ^{13}C-NMR data showed high similarity to the diketopiperazine **4**, indicating a roquefortine-like core structure. Furthermore, its ^{1}H-NMR spectrum revealed two protons at C-8 representing a single bond between C-8 and C-9, which is different from the double bond described for **13** (Supplemental Figure 1, Supplemental Table 1). Additionally, C-2 (δ_C = 146) in the ^{13}C-HMBC spectrum indicated a double bond between N-1 and C-2 which was supported by the chemical shift of N-1 (δ_N = 280) in the ^{15}N-HMBC spectrum (Supplemental Figure 2, Supplemental Table 1). As compound **13** was reported from various *Penicillium* species like *P. albocoremium* (Overy, et al., 2005), *P. glandicola* (Kozlovskii, et al., 1994) and *P. chrysogenum* (Vinokurova, et al., 2002) and proposed as a precursor of **7**, host and *roq* deletion strain chromatograms of *P. chrysogenum* were further analyzed for the presence of **13**. The chromatogram of the ion with *m/z* 404.1706, representing the protonated molecular $[M+H]^+$ with formula $C_{22}H_{22}N_5O_3$ of both compounds **6** and **13**, was extracted in a 5 ppm mass

accuracy window. However, no ion possibly corresponding to **13** could be found whereas **6** was observed at high concentration in the liquid media (Figure 3).

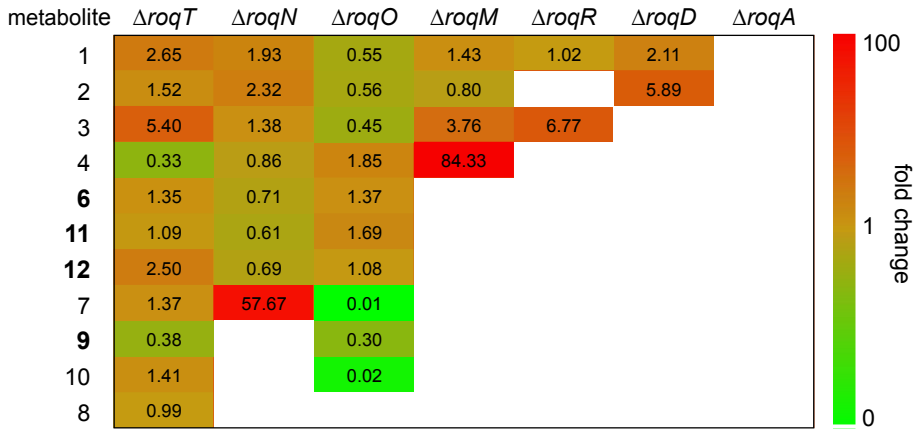


Figure 3. Fold change of the concentration of secondary metabolites from the roquefortine/meleagrin pathway in deletion strains compared to the host strain.
Numbers in table represent the internal standard corrected concentrations of secondary metabolites in supernatants of deletion strains obtained from HPLC-UV-MS compared to their concentration in the supernatant of the host strain *P. chrysogenum* DS54555. Cell coloring, representing the fold change, was performed on a logarithmic scale. Red colored cells indicate a concentration increase whereas green colored cells represent a decrease. Cells colored in white indicate a complete absence of the metabolite in the deletion sample. Novel metabolites of the roquefortine/meleagrin pathway are shown with bold numbers.

The absence of **13** in host and various *P. chrysogenum* strains lead to the conclusions that **13** is not produced by *P. chrysogenum* DS54555.
Compound **11** and **12** are novel compounds based on a roquefortine-like scaffold, thus named roquefortine M and roquefortine N. HR-ESI-MS of **11** (*m/z* 422.1814 $[M+H]^+$, calc. 422.1824) and **12** (*m/z* 440.1918 $[M+H]^+$, calc. 440.1928) established the molecular formula $C_{22}H_{23}N_5O_4$ and $C_{22}H_{25}N_5O_5$. Their chemical structure was determined using ^{13}C-HMBC, ^{15}N-HMBC and ^{1}H-NMR (Supplemental Figure 3 and 4, Supplemental Table 2) showing similar signals as observed for compound **6**, which indicates a similar chemical scaffold. However, the significant upfield shift of N-1 (from δ_N = 280 in **6** to δ_N = 185 in **11**) in the ^{15}N-HMBC spectra together with the chemical shift of C-2 (δ_C = 146 in **6**, δ_C = 172 in **11**) in the ^{13}C-NMR spectra of compound **11** shows that **11** contains a single bond between N-1 and C-2, with C-2 being a carbonylic carbon. In addition, a comparison between the ^{15}N-HMBC spectrum of compound **11** and **12** revealed that the amide-bond between N-14 and C-13 in **11** was hydrolyzed in **12** (δ_N = 32.0) yielding a primary amine and a carboxyl group. Both compounds commonly occur, together with compound **6**, in liquid cultures of *P. chrysogenum* host and *roqT*, *roqN* and *roqO* deletion strains (Figure 3). Their absence in the remaining deletion strain samples concludes the involvement of *roqA*, *roqR*, *roqD* and *roqM* in their biosynthesis.

*Structure elucidation and quantification of **9** and **10***
Compound **9** with molecular formula $C_{23}H_{25}N_5O_3$, established by HR-ESI-MS (*m/z*

420.2015 $[M+H]^+$, calc. 420.2030), was identified as roquefortine F, a metabolite solely reported from a deep-ocean sediment derived *Penicillium* species (Du, et al., 2009), using ^{1}H- and ^{13}C-NMR (Supplemental Figure 5, Supplemental Table 3). Its ^{1}H-NMR spectrum is very similar to the spectrum of **3** (Ali, et al., 2013), except a double bond between C-12 and C-15. Furthermore, the presence of C-26 (δ_C = 63.6) in the ^{13}C-NMR spectrum, next to a sharp OCH_3 peak (δ_H = 4.01) and a missing proton on N-1 in the ^{1}H-NMR spectrum fully agree with a methoxylated N-1 in compound **9**. This was supported by the absence of correlations with a carbon or proton in the HMBC spectrum. The concentration of **9**, particularly high in the host strain, was found to be reduced to approximately one third in the deletion strains of *roqT* and *roqO* and absent in the remaining deletion strains (Figure 3). This data suggest that *roqO* and *roqT* are the only two genes not involved in the biosynthesis of **9**. Compound **10** with molecular formula $C_{23}H_{25}N_5O_4$ (*m/z* 436.1967 $[M+H]^+$, calc. 436.1979) was identified as neoxaline, a metabolite previously isolated from *Aspergillus japonicas* Fg-551 (Hirano, et al., 1979) and *P. tulipae* (Overy, et al., 2006), by comparing retention time and MS/MS fragments to its commercially available standard (Supplemental Figure 6). While the concentration of **10** in host and *roqT* deletion strain is almost comparable, a 97% decrease was observed in the *roqO* deletion strain (Figure 3). In all remaining deletion strains, compound 10 could not be detected leading to the conclusion that all genes in the *roq* gene cluster, except *roqT*, are required for the synthesis of **10**.

*Chemical degradation of compound **6** leads to various products*

Nitrones, such as compound **6**, are not infinitely stable and degrade already at room temperature in aqueous solution as well as under acidic conditions by incorporation of water (Cashman, et al., 1999; Rodriguez, et al., 1999; Sun, et al., 2007). In order to determine the resulting degradation products, an aqueous solution of **6** was acidified and the resulting sample measured using HPLC-UV-MS (Figure 4). Next to a 50 % decrease of **6**, two highly abundant ions were observed in the treated sample corresponding to **11** and **12**. Additionally a third unidentified compound was found eluting at 17.33 minutes with a mass-to-charge ratio of 422.1823, representing the protonated molecule with the formula $C_{22}H_{24}N_5O_4$. These results demonstrate that **11**, **12** and an unidentified third compound are produced by degradation of the rather unstable compound **6**.

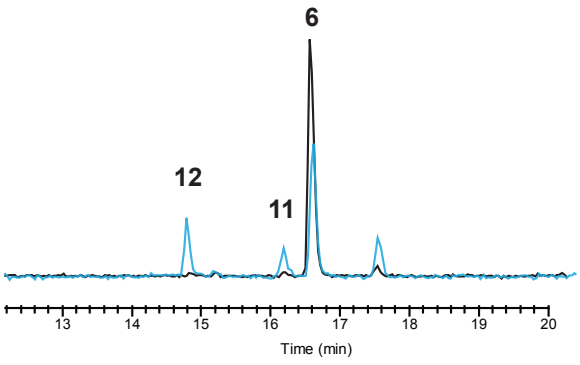


Figure 4. Chemical degradation of compound **6** leads to various products.
Total ion chromatogram of pure (black) and degraded compound **6** (blue) after addition of formic acid measured on HPLC-MS. Acid induced degradation leads to the formation of **11** and **12** next to an unidentified compound eluting at 17.33 minutes with $[M+H]^+$= 422.1823 and elemental composition $C_{22}H_{23}N_5O_4$.

Discussion

Here, we present new insight into the complex biosynthesis of secondary metabolites from the roquefortine/meleagrin pathway. Five novel metabolites were found to originate from the *roq* gene cluster, obtained from comparative metabolites profiling of the host strain and various deletion strains in combination with NMR and MS based structure elucidation. As all five metabolites are produced in a late stage of the pathway, no changes were observed for the biosynthesis of upstream metabolites **1** – **4**, which starts with RoqA taking L-histidine and L-tryptophan as substrates and producing compound **1**. Based on the highly significant accumulation of **4** in the *roqM* deletion strain and the absence of all downstream metabolites **6** – **12** (Figure 3), it can be concluded that *roqM*, encoding a flavin-dependent MAK 1-monooxygenase like protein, is involved in the conversion of **4** into **6**, a novel compound containing an unusual nitrone moiety. Nitrones are widely known due to their free radical trapping properties and their potential application as therapeutics in age related diseases (Floyd, et al., 2008) like cancer (Floyd, et al., 2011) and ischaemic stroke (Maples, et al., 2004). As the chemical scaffold of compound **6** is closely related to the roquefortine group it was named roquefortine L. Flavin-containing monooxygenases are commonly known to consecutively oxidize drugs and xenobiotics containing a soft-nucleophile, such as nitrogen or sulfur (Krueger and Williams, 2005). In case of secondary amines, flavin-containing monooxygenases consecutively oxidize the nitrogen, leading to the production of hydroxylamines and nitrones (Cashman, et al., 1999; Rodriguez, et al., 1999; Sun, et al., 2007). A similar mechanism for the synthesis of the nitrone containing compound **6** is very likely, starting with the oxidation of the secondary amine in the indole part of **4**, yielding the hydroxylated intermediate **5** (Figure 5). Further oxidation on the same

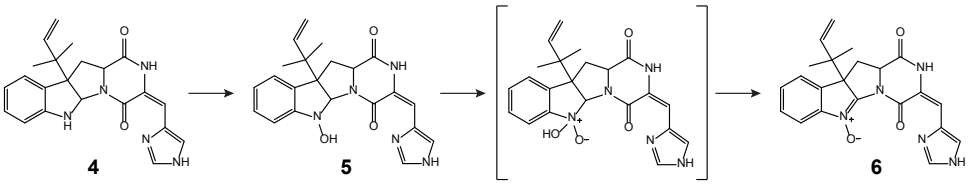


Figure 5. Proposed biosynthesis of **6** by RoqM.

nitrogen produces an unstable N,N-dihydroxylated species which is followed by the loss of water, producing eventually compound **6**. However, nitrones are not indefinitely stable and easily degrade at room temperature in aqueous solutions (Cashman, et al., 1999; Rodriguez, et al., 1999; Sun, et al., 2007). Under acidic conditions compound **6** decomposes by a consecutive incorporation of water leading, among others, to the production of compound **11** and **12** (Figure 4 and 6). This decomposition was also observed in NMR experiments after extended storage of a solution of 6 at room temperature. These results suggest that the presence of **11** and **12** in liquid cultures of *P. chrysogenum* can be attributed to a chemical degradation of **6**. Compound **6**, with the formula $C_{22}H_{21}N_5O_3$, is represented by an ion with a mass-to-charge ratio of 404.1706 and eluting at 16.8 minutes. The exact same ion was

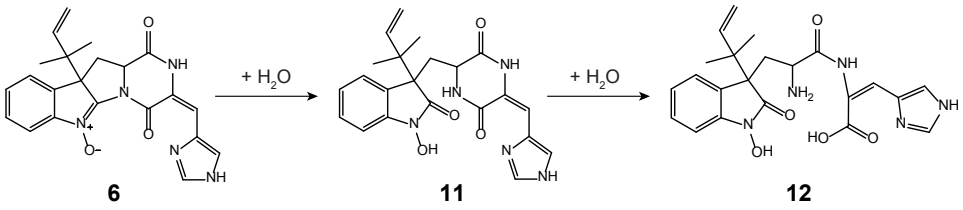


Figure 6. Degradation of **6** yielding **11** and **12** by consecutive incorporation of water.

previously tentatively identified as compound **13** (Ali, et al., 2013) as its elemental composition and parts of the structure indicated consistency with this compound. However, further structure elucidation using various NMR experiments confirmed the structure of **6** instead. This was surprising as compound **13**, a proposed key-intermediate in the biosynthesis of downstream metabolites like **7**, **8** and **10** was previously tentatively identified in different *Penicillium* cultures (Overy, et al., 2005; Vinokurova, et al., 2002). By using a comparable instrumental set-up with a similar chromatographic separation method, host and deletion strains of DS54555 were screened for production of **13**. Nevertheless, neither **13** nor corresponding degradation products could be detected, whereas **6** was found at high concentrations leading to the conclusion that **13** is not produced by *P. chrysogenum*. This is remarkable as **13** was expected as single precursor of **7**, modified by RoqO (Garcia-Estrada, et al., 2011). In addition, a deletion of *roqO* resulted in an up to 98 % decrease of **7** whereas the levels of upstream metabolites remained nearly unchanged, indicating that RoqO is indeed involved in the synthesis of **7**. Due to the general absence of **13** in the *P. chrysogenum* derived samples, compound **7** has to originate from a different biosynthetic route than the previously reported oxidation of **13** on its indole nitrogen (Ali, et al., 2013; Garcia-Estrada, et al., 2011; Overy, et al., 2005). RoqO, encoding a p450 monooxygenase, closely resembles FtmG (64% identity, 79% similarity at the amino acid level) a cytochrome p450 monooxygenase catalyzing the hydroxylation of fumitremorgin C to dihydroxy-fumitremorgin C (Kato, et al., 2009), compounds that are structurally similar to the roquefortine derivatives. A possible deduced biosynthesis of **7** involves the hydroxylation of **6** on C-9 by RoqO similar to the oxidation of fumitremorgin C by FtmG. Subsequent cleavage of the bond between C-9 and N-14 followed by the development of a bond between C-2 and N-11 is postulated to yield ultimately **7**. These results, together with the general absence of **13** in *P. chrysogenum* lead to the conclusion that **7** and **8** are produced via a different biosynthesis in *P. chrysogenum* than in other *Penicillium* strains like *P. tulipae* (Overy, et al., 2005), for which a tentatively identified **13** was reported as intermediate.

The deletion of *roqN* resulted in an accumulation of **7** in the liquid medium whereas metabolites **8**, **9** and **10** were absent (Figure 3). RoqN, a methyltransferase, was previously recognized to catalyze the addition of a methyl group on the hydroxylated nitrogen of **7** producing **8** (Ali, et al., 2013; Garcia-Estrada, et al., 2011). As **9** contains a methylated hydroxylamine group in the same position as **8**, the hydroxylamine containing compound **5**, which differs only in a methyl-group, is proposed as its direct precursor with RoqN catalyzing the methyl addition to yield **9**. These results

reveal a further branching of the roquefortine/meleagrin pathway with compounds **6** and **9** being products of **5**. In addition, they support the presence of **5**, which was proposed based on its involvement in the biosynthesis of **6**.
Compound **10** was previously proposed as direct product of **8** by enzymatic hydrogenation (Overy, et al., 2005). However, BLAST analysis did not reveal an enzyme in the roquefortine/meleagrin pathway, which is able to perform that reaction (Ali, et al., 2013; Garcia-Estrada, et al., 2011). Moreover, a 53 times higher concentration of **8** compared to **10** in the host strain, but the absence of **8** in the *roqO* deletion strain with **10** still being present, leads to the conclusion that **8** is not a precursor of **10** (Figure 3). In contrast, due to the high concentration of **9** in the *ΔroqO* strain and its roquefortine-like structure (roquefortine scaffold with a methoxygroup on N-1), compound **9** is proposed as direct precursor of **10** with RoqO catalyzing this reaction, similar to the synthesis of **7** from **6**. These results suggest that RoqO is involved in the reactions from **6** into **7** and from **9** into **10** by oxidizing and subsequently converting a roquefortine scaffold into a glandicoline like structure (Figure 1B).
In conclusion, these results extend the additional branch of compound **9** leading to the final product **10**. Unspecificity, already observed for RoqR and RoqD (Ali, et al., 2013), could now also be observed for RoqO and RoqN leading to a complex degree of branching in the pathway and a wide palette of compounds. Several of the new compounds identified in the current study were found to be equipped with interesting biological activities. Roquefortine F, previously reported from a deep ocean sediment derived fungus *Penicillium* sp., shows moderate cytotoxicity against various tumor cell lines (Du, et al., 2009). Neoxaline, which was first isolated from *A. japonicas* Fg-551, stimulates the central nervous system in mice (Hirano, et al., 1979) and inhibits cell proliferation (Koizumi, et al., 2004). Furthermore, it was found to induce cell cycle arrest at the G2/M phase in Jurkat cells (inhibition of tubulin polymerization) (Koizumi, et al., 2004). Here, the novel metabolites roquefortine L, roquefortine M and roquefortine N are added to the palette of potential cytotoxic compounds, which demonstrates the potential of engineered industrial *P. chrysogenum* strains to produce novel bioactive compounds with unusual chemical scaffolds.

Acknowledgements

We would like to thank Drs. H. Menke, W. Heijne and H. Roubos from the DSM Biotechology Centre for making the DNA microarray data available.

References

Ali, H., Ries, M.I., Nijland, J.G., Lankhorst, P.P., Hankemeier, T., Bovenberg, R.A., Vreeken, R.J., and Driessen, A.J. (2013). A Branched Biosynthetic Pathway Is Involved in Production of Roquefortine and Related Compounds in Penicillium chrysogenum. PLoS One 8, e65328.

Cashman, J.R., Xiong, Y.N., Xu, L., and Janowsky, A. (1999). N-oxygenation of amphetamine and methamphetamine by the human flavin-containing monooxygenase (form 3): role in bioactivation and detoxication. J Pharmacol Exp Ther 288, 1251-1260.

Clark, B., Capon, R.J., Lacey, E., Tennant, S., and Gill, J.H. (2005). Roquefortine E, a diketopiperazine from an Australian isolate of Gymnoascus reessii. J Nat Prod 68, 1661-1664.

Du, L., Feng, T., Zhao, B., Li, D., Cai, S., Zhu, T., Wang, F., Xiao, X., and Gu, Q. (2010). Alkaloids from a deep ocean sediment-derived fungus Penicillium sp. and their antitumor activities. J Antibiot (Tokyo) 63, 165-170.

Du, L., Li, D., Zhu, T., Cai, S., Wang, F., Xiao, X., and Gu, Q. (2009). New alkaloids and diterpenes from a deep ocean sediment derived fungus Penicillium sp. Tetrahedron 65, 1033-1039.

Floyd, R.A., Chandru, H.K., He, T., and Towner, R. (2011). Anti-cancer activity of nitrones and observations on mechanism of action. Anticancer Agents Med Chem 11, 373-379.

Floyd, R.A., Kopke, R.D., Choi, C.H., Foster, S.B., Doblas, S., and Towner, R.A. (2008). Nitrones as therapeutics. Free Radic Biol Med 45, 1361-1374.

Garcia-Estrada, C., Ullan, R.V., Albillos, S.M., Fernandez-Bodega, M.A., Durek, P., von Dohren, H., and Martin, J.F. (2011). A single cluster of coregulated genes encodes the biosynthesis of the mycotoxins roquefortine C and meleagrin in Penicillium chrysogenum. Chemistry & biology 18, 1499-1512.

Hirano, A., Iwai, Y., Masuma, R., Tei, K., and Omura, S. (1979). Neoxaline, a new alkaloid produced by Aspergillus japonicus. Production, isolation and properties. J Antibiot (Tokyo) 32, 781-785.

Kato, N., Suzuki, H., Takagi, H., Asami, Y., Kakeya, H., Uramoto, M., Usui, T., Takahashi, S., Sugimoto, Y., and Osada, H. (2009). Identification of cytochrome P450s required for fumitremorgin biosynthesis in Aspergillus fumigatus. ChemBiochem 10, 920-928.

Koizumi, Y., Arai, M., Tomoda, H., and Omura, S. (2004). Oxaline, a fungal alkaloid, arrests the cell cycle in M phase by inhibition of tubulin polymerization. Biochim Biophys Acta 1693, 47-55.

Koolen, H.H., Soares, E.R., Silva, F.M., Souza, A.Q., Medeiros, L.S., Filho, E.R., Almeida, R.A., Ribeiro, I.A., Pessoa Cdo, O., Morais, M.O., et al. (2012). An antimicrobial diketopiperazine alkaloid and co-metabolites from an endophytic strain of Gliocladium isolated from Strychnos cf. toxifera. Nat Prod Res 26, 2013-2019.

Kozlovskii, A.G., Vinokurova, N.G., Reshetilova, T.A., Sakharovskii, V.G., Baskunov, B.P., and Seleznev, S.G. (1994). New metabolites of Penicillium glandicola var. glandicola - glandicoline A and glandicoline B. Appl Biochem Microbiol 30, 334-337.

Krueger, S.K., and Williams, D.E. (2005). Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. Pharmacol Ther 106, 357-387.

Maples, K.R., Green, A.R., and Floyd, R.A. (2004). Nitrone-related therapeutics: potential of NXY-059 for the treatment of acute ischaemic stroke. CNS Drugs 18, 1071-1084.

Overy, D.P., Nielsen, K.F., and Smedsgaard, J. (2005). Roquefortine/oxaline biosynthesis pathway metabolites in Penicillium ser. Corymbifera: in planta production and implications for competitive fitness. J Chem Ecol 31, 2373-2390.

Overy, D.P., Phipps, R.K., Frydenvang, K., and Larsen, T.O. (2006). epi-Neoxaline, a chemotaxonomic marker for Penicillium tulipae. Biochem Syst Ecol 34, 345-348.

Rodriguez, R.J., Proteau, P.J., Marquez, B.L., Hetherington, C.L., Buckholz, C.J., and O'Connell, K.L. (1999). Flavin-containing monooxygenase-mediated metabolism of N-deacetyl ketoconazole by rat hepatic microsomes. Drug Metab Dispos 27, 880-886.

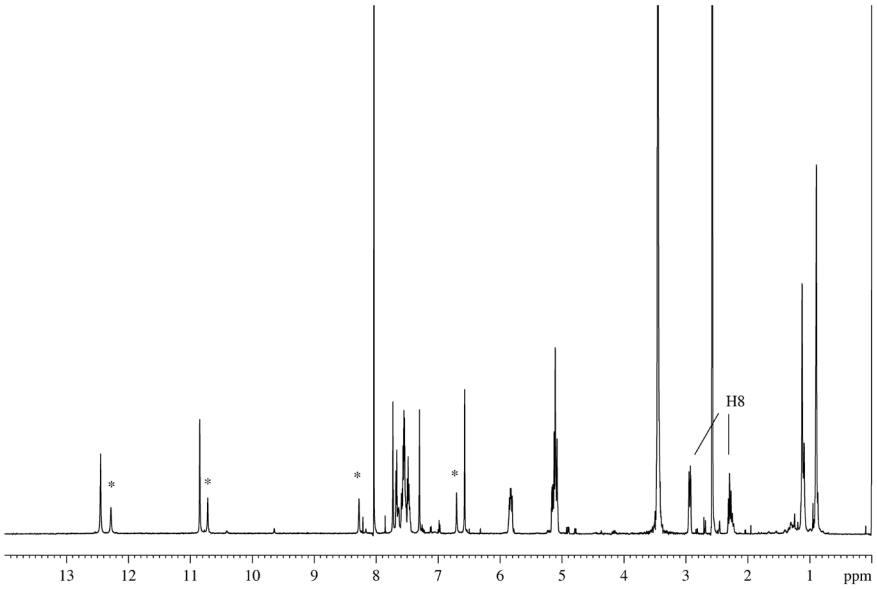
Scott, P.M., Merrien, M.A., and Polonsky, J. (1976). Roquefortine and iso-fumigaclavine A, metabolites from Penicillium roqueforti. Experientia 32, 140-142.

Sun, H., Ehlhardt, W.J., Kulanthaivel, P., Lanza, D.L., Reilly, C.A., and Yost, G.S. (2007). Dehydrogenation of indoline by cytochrome P450 enzymes: a novel "aromatase" process. J Pharmacol Exp Ther 322, 843-851.

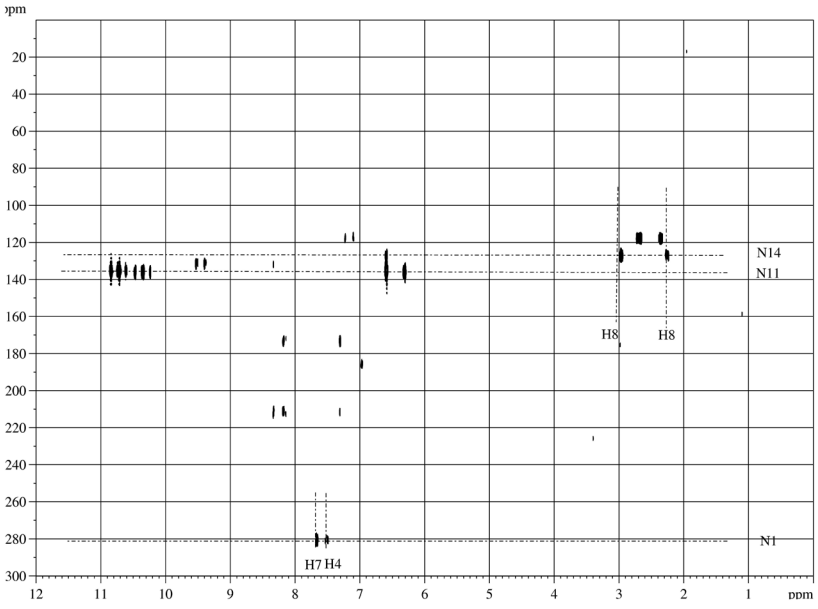
Vinokurova, N.G., Boichenko, L.V., and Arinbasarov, M.U. (2002). Production of Alkaloids by Fungi of the Genus Penicillium Grown on Wheat Grain. Appl Biochem Microbiol 39, 403-406.

Weber, S.S., Polli, F., Boer, R., Bovenberg, R.A., and Driessen, A.J. (2012). Increased penicillin production in Penicillium chrysogenum production strains via balanced overexpression of isopenicillin N acyltransferase. Appl Environ Microbiol 78, 7107-7113.

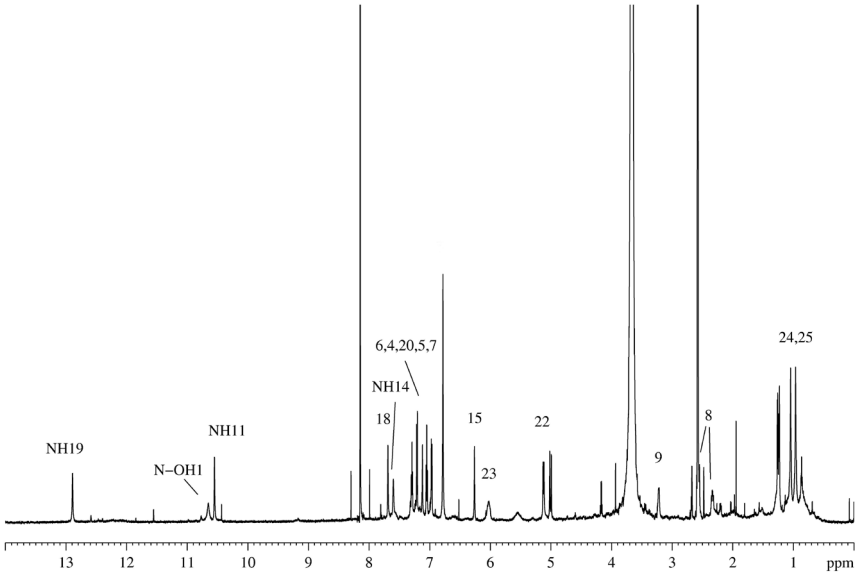
Supplemental Information



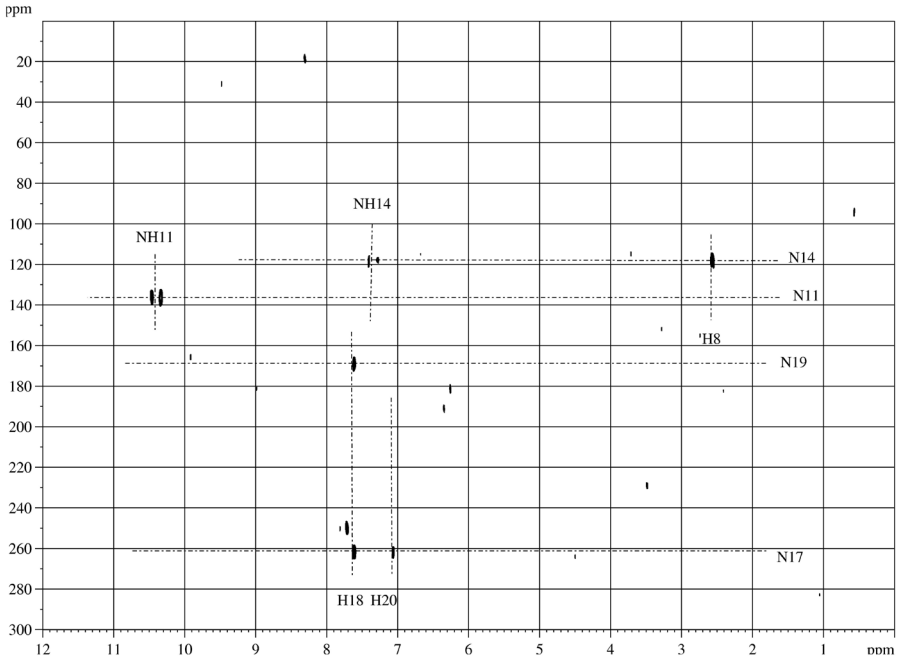
Supplemental Figure 1. ^{1}H-NMR spectrum of roquefortine L (**6**) in DMSO/$CDCl_3$ acquired at 300K on a 600 MHz spectrometer. Small additional peaks labeled with a star correspond to a second conformation of **6**. Assignments can be found in Supplemental Table 1.



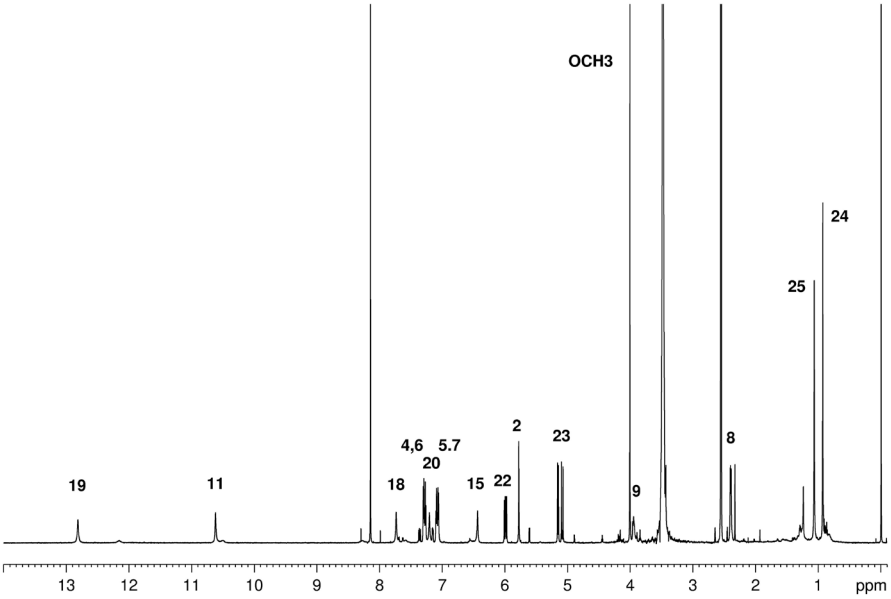
Supplemental Figure 2. ^{15}N-HMBC spectrum of compound **6** acquired at 290 K. Small additional peaks correspond to compound 11 which was produced by slow degradation of **6**. Further assignments can be found in Supplemental Figure 1.



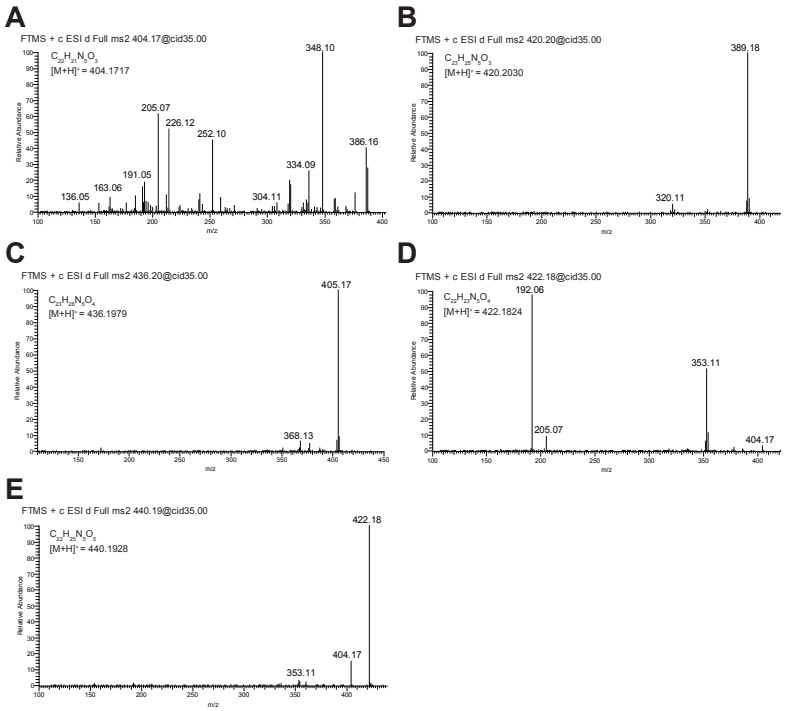
Supplemental Figure 3. ^{1}H-NMR spectrum of roquefortine M (**11**) acquired at 250 K.



Supplemental Figure 4. ^{15}N-HMBC spectrum of roquefortine M (**11**) acquired at 270 K.

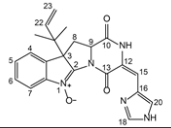


Supplemental Figure 5. ^{1}H-NMR spectrum of roquefortine F (**9**) acquired at 280 K.



Supplemental Figure 6. HPLC-MS/MS fragmentation spectra including chemical formula and calculated exact mass of metabolites. Metabolites shown are protonated compound **6** (roquefortine L, A), **9** (roquefortine F, B), **10** (neoxaline, C), **11** (roquefortine M, D) and **12** (roquefortine N, E). Spectra were acquired at a LTQ-FT-MS at 35% normalized collision energy in positive ion mode.

roquefortine L (**6**)

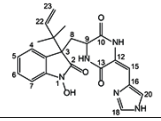


	290 K	280 K	290 K
	δ_H	δ_C	δ_N
1-NO	-		280
2		147.6	
3		58.2	
3a		137.9	
4	7.49	124.3	
5	7.44	129.0	
6	7.52	128.6	
7	7.56	114.9	
7a		148.5	
8	2.24, 2.95	24.5	
9	5.06	63.6	
10		164.9	
11-NH	10.77		136
12		122.2	
13		156.7	
14-N			127
15	6.57	112.2	
16		124.3	
17-N			*
18	7.76	136.3	
19-NH	12.42**		*
20	7.49	132.1	
21		43.1	
22	5.79	141.5	
23	5.06, 5.09	116.2	
24	0.86	22.4	
25	1.09	22.3	

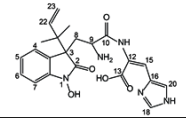
(*) not observed
(**) from spectrum at 280 K

Supplemental Table 1
Chemical shifts of ^{1}H, ^{13}C and ^{15}N-NMR of roquefortine L (**6**) (δ in ppm).

roquefortine M (**11**)



roquefortine N (**12**)

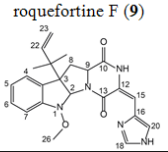


	270K	300K	270K		280K	280K	300K
	δ_H	δ_C	δ_N		δ_H	δ_C	δ_N
1-NOH	10.56		185***		10.64**		*
2		172.6				173.1	
3		65.9				54.5	
3a		124.7				124.7	
4	7.20	125.1			7.18	125.0	
5	7.02	121.6			6.99	121.1	
6	7.27	128.2			7.22	127.6	
7	6.95	106.8			6.93	106.4	
7a		142.9				142.7	
8	2.32, 2.60	35.6			2.09, 2.58	35.0	
9	3.20	52.3			2.74	52.6	
10		165.1				*	
11-NH	10.44		137		10.08**		*
12		121.0				*	
13		160.1				*	
14-NH	7.38		118	14-NH_2	6.57**		30
15	6.26	108.7			8.02	107.1	
16		125.3				127.7	
17-N			262				*
18	7.65	136.4			8.00	133.4	
19-NH	12.87		169		*		196
20	7.09	133.8			7.18	127.6	
21		41.9				41.7	
22	6.00	142.3			6.03	142.5	
23	4.98, 5.09	113.9			4.98, 5.09	113.4	
24	1.04	21.7			1.07	21.4	
25	0.96	21.3			0.98	21.2	

(*) not observed
(**) tentative assignment
(***) from spectrum at 290 K

Supplemental Table 2
Chemical shifts of ^{1}H, ^{13}C and ^{15}N-NMR of roquefortine M (**11**) and roquefortine N (**12**) (δ in ppm).

roquefortine F (**9**)



	280K	280K
	δ_H	δ_C
2	5.78	84.9
3		60.5
3a		130.5
4	7.30	124.5
5	7.09	124.4
6	7.27	128.8
7	7.06	116.0
7a		150.9
8	2.40	37.7
9	3.96	57.4
10		164.5
11-NH	10.63	
12		*
13		157.4
14-N14		
15	6.44	109.1
16		*
18	7.75	136.9
19-NH	12.82	
20	7.21	134.2
21		40.5
22	5.99	143.1
23	5.09, 5.16	114.6
24	0.94	23.2
25	1.07	22.4
26	4.01	63.6

(*) not observed

Supplemental Table 3
Chemical shifts of ^{1}H and ^{13}C-NMR of roquefortine F (**9**) (δ in ppm).

Chapter 4

A single unspecific non-linear NRPS is involved in the synthesis of cyclic tetrapeptides in *Penicillium chrysogenum*

Based on

Hazrat Ali*, Marco I. Ries*, Peter P. Lankhorst, Rob A.M. van der Hoeven, Olaf L. Schouten, Marek Noga, Thomas Hankemeier, Noël N.M.E. van Peij, Roel A.L. Bovenberg, Rob J. Vreeken, Arnold J.M. Driessen

A single unspecific non-linear NRPS is involved in the synthesis of cyclic tetrapeptides in Penicillium chrysogenum

Submitted for publication

** these authors contributed equally*

Abstract

The filamentous fungus *Penicillium chrysogenum* harbors an astonishing variety of nonribosomal peptide synthetase genes, which encode proteins known to produce complex bioactive metabolites from simple building blocks. Here we report a novel non-linear tetra-modular nonribosomal peptide synthetase (NRPS) with a trans-acting adenylation domain and with a distinctive non-specificity of all involved adenylation domains towards their respective substrates. By deleting the putative gene in combination with comparative metabolite profiling various unique cyclic and linear tetrapeptides were identified which were associated with this NRPS. In combination with substrate predictions for each module, we propose a detailed mechanism of the 'trans-acting' adenylation domain.

Introduction

Fungal non-ribosomal peptides contribute a large variety of secondary metabolites with remarkable properties such as antibacterial, antifungal, antiparasitic, anticancer and immunosuppressive activities. These metabolites are produced by large, multifunctional protein complexes, called nonribosomal peptide synthetases (NRPS) that catalyze the stepwise condensation of simple amino acid building blocks to complex molecules. NRPSs have a modular organization, with each module responsible for one discrete chain-elongation step. Every single module can be subdivided into domains that carry all essential information for recognition, activation and modification of one substrate. At a minimum, a typical NRPS module consists of an adenylation (A) domain, responsible for amino acid activation, a thiolation domain, also known as peptidyl carrier protein (PCP), which binds the activated amino acid and a condensation (C) domain that catalyzes peptide-bond formation. The common arrangements of these domains follow a $(C\text{-}A\text{-}PCP)_n$ organization. Additionally, a variety of optional domains have been described such as methyltransferase (MT) and epimerization (E) domains (Schwarzer and Marahiel, 2003).
The number of modules and their domain organization within NRPS enzymes controls the structures of the final product(s) (Grunewald and Marahiel, 2006; Mootz et al., 2002; Schwarzer et al., 2003). Thus, the order of modules usually corresponds to the sequence of amino acids in the peptide. Many NRPS systems adhere to this mechanistic paradigm, which is often referred to as the "co-linearity rule" (Fischbach and Walsh, 2006). Also exceptions to this rule have been discovered, including iterative NRPSs, which incorporate multiple residues of the same amino acid iteratively into the peptide structure and the so called nonlinear NRPSs, which deviate completely from the standard domain organization leading to unexpected products (Mootz et al., 2002; Shaw-Reid et al., 1999).
The impact of non-ribosomal peptide metabolites on the quality of human life raised the interest of pharmaceutical industries to invest in identification, engineering and heterologous expression of NRPS genes and pathways to ensure the rational production of novel compounds (Stevens et al., 2006; Wyatt et al., 2012; Zhang et al., 2013). To understand the basic mechanisms of the biosynthesis of these complex NRPSs, detailed studies have been performed during the past few decades. These included the structural analysis of adenylation domains, mutational analysis of substrate specificity of these modules, the fusion of unrelated modules to produce new products and the identification of helper proteins for optimal activation of adenylation domains (Baltz, 2011; Conti et al., 1997; Doekel et al., 2008). Although this has led to detailed insights into catalytic mechanisms, so far a structure of a complete NRPS is lacking that would reveal how modules cooperate to facilitate product formation. The availability of genome sequencing data and sophisticated bioinformatics analysis of various fungi revealed the presence of many NRPS genes that have not been associated with known secondary metabolites (Khaldi et al., 2010; Pel et al., 2007; van den Berg et al., 2008). Moreover, most of these genes are not expressed when the fungi are grown under laboratory conditions, implying that many more secondary metabolites await discovery.
The filamentous fungus *Penicillium chrysogenum* is well known for the production of the antibiotic penicillin G which is synthesized by the tri-modular NRPS δ-(L-α-

aminoadipyl)-L-cysteinyl-D-valine synthetase. In addition, other NRPS derived secondary metabolites like the roquefortine toxins and meleagrin have been reported

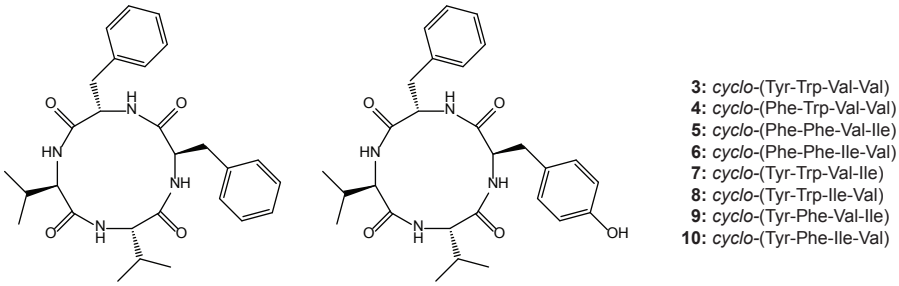


Figure 1. Identified secondary metabolites
Structures of cyclic tetrapeptides identified in *P. chrysogenum.*

from *P. chrysogenum* (Ali et al., 2013; Frisvad et al., 2004; Garcia-Estrada et al., 2011). Here, we describe the identification and structural characterization of cyclic tetrapeptides (Figure 1) and the discovery of a tetra-modular NRPS with an unusual domain organization and limited specificity of its adenylation domains. It is proposed to term this NRPS PcHcpA based on the produced Hydrophobic cyclic peptides.

Experimental procedures

Chemicals
AQC reagent and borate buffer were obtained as part of AccQ Tag Reagent Kit from Waters (Waters, Milford, MA, USA). *cyclo*-(*d*-Tyr-*l*-Phe-*d*-Val-*l*-Val) was obtained from Celtek Peptides (Nashville, TN).

Host strains, media, grown condition and plasmid construction
Deletion of the *PcHcpA* gene was carried out in *P. chrysogenum* strain DS54555, which lacks penicillin cluster genes and the *Ku70* gene (Ali et al., 2013). This strain was kindly provided by the DSM Biotechnology Center (Delft, Netherlands). Deletion plasmids were constructed by amplifying the flanking regions of the targeted gene with the Multisite Gateway®Three-Fragment Vector Construction Kit (Invitrogen). *Escherichia coli* (DH5α) was used as host strain for high frequency transformation and plasmid DNA amplification (Sambrook et al., 1989). All the strains were grown on YGG-medium for protoplasts formation and transformation (Kovalchuk et al., 2012). Both mutant and host strains of *P. chrysogenum* were grown on SMP medium as described previously (Ali et al., 2013).

Transformation procedure
Deletion plasmids were transformed to the protoplasts of *P. chrysogenum* DS54555 (Alvarez et al., 1987). The phleomycin resistance gene was used as selection marker for the deletion of the *PcHcpA* gene (Kolar et al., 1988; Kovalchuk et al., 2012).

Genomic DNA extraction, total RNA extraction, cDNA amplification and qPCR analysis

Genomic DNA (gDNA) was isolated after 96 hours of growth on SMP medium using the modified yeast gDNA isolation protocol (Harju et al., 2004) in which the fungal mycelium is broken in a FastPrep FP120 system (Qbiogene). Isolated gDNA was measured using a NanoDrop ND-1000. gDNA of the host and various deletion strains was isolated using the E.Z.N.A. Fungal DNA kit (Omega Bio-tek). Total RNA of the host strain was isolated after 48 hours of growth in SMP medium for the first time and then with the interval of 24 hours up to 216 hours of growth using the Trizol reagent (Invitrogen), with additional DNase treatment using the Turbo DNA-free kit (Ambion). Total RNA was measured with the NanoDrop ND-1000 and a concentration of 500 ng per cDNA reaction was used. cDNA was synthesized using the iScript cDNA synthesis kit (Bio-Rad) in a 10 µl end volume. The primers used to analyze the expression of the PcHcpA gene were designed around an intron to avoid amplification on gDNA (Table S3). For expression analyses, the γ-actin gene was used as a control for normalization. A negative reverse transcriptase (RT) control was used to determine the gDNA contamination in isolated total RNA. The expression levels were determined as described (Ali et al., 2013).

Southern blotting

Southern blotting was carried out by digesting gDNA (5 µg) with the indicated restriction enzymes. Digested DNA fragments were separated on a 0.8 % agarose gel, blotted onto a Zeta-Probe membrane (Biorad) as described earlier (Nijland et al., 2008), and hybridized with the indicated DIG labeled probes.

Metabolite profiling

Host and deletion strains of *P. chrysogenum* strains used for gene assignments were grown in quintuplicate according to the procedure described above. Samples for acquisition of the metabolite profiles from the growth curves were from five replicates. Metabolite profiling was carried out with modifications as described earlier (Ali et al., 2013). Briefly, 4 µl of internal standard mixture (855 nmol/mL ranitidine, 657 nmol/mL reserpine and 114 nmol/mL ampicillin) was added to 100 µl fermentation broth followed by the addition of 400 µl methanol for protein precipitation. The samples were vortexed and spun down at 14,000 g for 10 minutes. 300 µL supernatant was evaporated for 30 minutes in a speedvac (Thermo Scientific, San Jose, CA) and re-dissolved in 100 µL water. LC-UV-MS analysis was performed on an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled in-line to a Surveyor PDA detector (Thermo Scientific, San Jose, CA) and LTQ-FT mass spectrometer (Thermo Scientific, San Jose, CA) using electrospray ionization and operated in a scan range between *m/z* 110 and *m/z* 2000 in positive/negative ion switching mode. Separation was performed on a Waters Atlantis T3 column (2.1 x 100 mm, 3 µm) (Waters, Milford, MA) starting with 98 % of solvent A (1 % acetonitrile and 0.1 % formic acid in water) and 2 % solvent B (1 % water and 0.1 % formic acid in acetonitrile) for 1.5 minutes at a flow rate of 300 µL/min. 40 % B were reached after 22 minutes and 100 % B at 25 minutes. The column was flushed with 100 % B and re-equilibrated to initial conditions. Peak detection and integration were performed using an in-house

tool followed by statistical tests to discover significant different features. Finally, discovered features were integrated using LCquan v.26 (Thermo Scientific, San Jose, CA). The non-related non-endogenous compound reserpine was used as internal standard.

Identification of cyclic tetrapeptides
The identity of cyclic tetrapeptides was determined using samples from liquid cultures of *P. chrysogenum* and a crude precipitate containing primarily **1** and **2** next to various minor abundant cyclic tetrapeptides. LC-MS^n experiments for the determination of consecutive amino acid losses were performed according the metabolite profiling section with normalized collision energies of 35 %, an isolation width of *m/z* 1 and an activation Q of 0.30. NMR spectra were recorded on a Bruker Avance III 700 MHz NMR spectrometer (Bruker, Billerica, MA), equipped with a 5 mm TCI probe. 2 mg of each sample was dissolved in 0.6 mL anhydrous DMSO. NMR spectra were acquired at 340 K.

Identification of linear tetrapeptides
Linear tetrapeptides were identified according their multiple-stage fragmentation after AQC derivatization (Noga et al., 2012). Methanol (400 µL) was added to an aliquot of 100 µL fermentation broth for protein precipitation. Samples were vortexed for 10 minutes, spun down for 10 minutes and 300 µL of the supernatant was evaporated to dryness in a speedvac (Thermo Scientific, San Jose, CA). Derivatization was done according to the supplier's procedure by re-dissolving the sample in 40 µL water, 40 µL borate buffer (pH 8.5) and 20 µL AQC solution. The mixture was vortexed for 10 minutes and heated for 10 minutes at 55°C.
LC-MS^n experiments were conducted on an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled to a LTQ-FT mass spectrometer (Thermo Scientific, San Jose, CA) using electrospray ionization. Separation was performed on a Waters Atlantis T3 column (2.1 x 100 mm, 3 µm) (Waters, Milford, MA) starting with 72 % of solvent A (1 % acetonitrile and 0.1 % formic acid in water) and 28 % of solvent B (1 % water and 0.1 % formic acid in acetonitrile) for 1.5 minutes at a flow rate of 300 µL/min. After 8 minutes the gradient reached 60 % of solvent B. Subsequently, the column was flushed with 100 % B before it was re-equilibrated to initial conditions. The peptide sequences were elucidated using multiple-stage collision-induced dissociation (CID) of the protonated molecule following consecutive cleavages of amino acids residues, starting from the C-terminus of the derivatized linear tetrapeptide. CID was performed with normalized collision energies of 35 %, an isolation width of *m/z* 1 and an activation Q of 0.30.

Results

Bioinformatic analysis of a tetrapeptide NRPS
Genome sequencing revealed that *P. chrysogenum* encodes 11 NRPS genes (van den Berg et al., 2008). Microarray expression analysis under glucose-limited chemostat culture conditions as well as quantitative PCR under shake flask culture condition showed that Pc16g04690 (*PcHcpA*) is highly expressed (Figure 2A) (van den Berg et

al., 2008). The *PcHcpA* gene encodes a large multimodular non-ribosomal peptide synthetase enzyme (Figure 3) with 6064 amino acids and a calculated molecular mass of about 670 kDa. PcHcpA, which shows 54% sequence identity to the orthologous An08g02310 (AnHcpA) in *A. niger*, has the domain architecture A_1-PCP_1-E-C_2-A_4-A_2-PCP_2-C_3-A_3-PCP_3-E-C_4-PCP_4-C-PCP (A = adenylation, C = condensation, PCP = thiolation, and E = epimerization) (Figure 3A) (van den Berg et al., 2008). A similar domain architecture was deduced for the orthologous AnHcpA protein of *A. niger* except for an insertion of a 177 amino acid long sequence between the adenylation domain A_4 and A_2 which shows homology to conserved motifs of an incomplete condensation domain (C_o) (Figure 3A). To predict the substrate specificity of the four adenylation domains of PcHcpA and AnHcpA, NRPSPredictor2 was used (Rottig et al., 2011). This program extracted the active site amino acid motifs DAACVAGVAK as the signature sequence for the A_1 domain in PcHcpA and DAVIIAAVAK in AnHcpA, which have similarity with the signature sequence of the adenylation domain of a bacitracine-producing NRPS that activates phenylalanine as a substrate. The signature sequences for the A_4 domain (DAVSAGVAAK in PcHcpA and DMQSAWFICK in AnHcpA) shows homology with the valine-activating adenylation domain of the

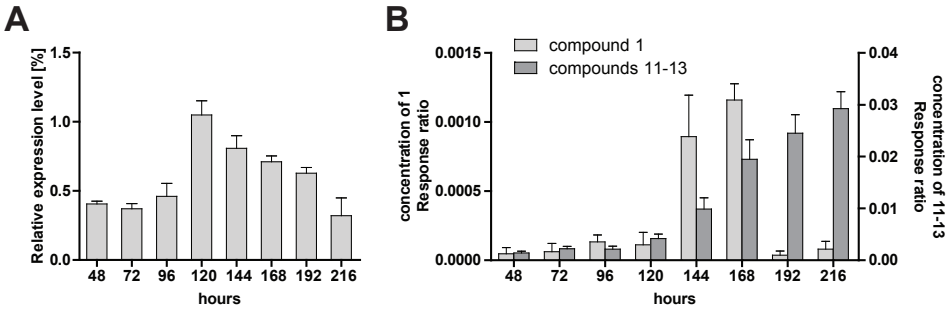


Figure 2. Correlation between the expression level of *PcHcpA* and metabolite formation
A: Time dependent expression level of *PcHcpA* as determined by quantitative RT-PCR.
B: Internal standard corrected concentration of the cyclic peptide **1** and its degradation products **11**-**13** present in the growth media. The concentration of peptides was determined by HPLC-UV-MS.

gramicidin synthetase, whereas the A_2 (DAMTVGGVFK for PcHcpA and DVLSTGAICK for AnHcpA) and A_3 (DAMFVGGVFK for PcHcpA and DAMFVGGIFK for AnHcpA) domains have predicted specificities towards phenylalanine and valine, respectively. The overall architecture of both synthetases is unusual, as the A_2 and A_4 domains occur adjacent to each other, flanked by a single C and PCP domain in a C_2-A_4-A_2-PCP_2 pattern. On the other hand, an incomplete module (C_4-PCP_4) without an adjacent A domain is found at the N-termini of these NRPSs (Figure 3A).

Genetic deletion of the tetrapeptide NRPS and secondary metabolite identification

In order to identify the secondary metabolites synthetized by PcHcpA, the corresponding gene was deleted and comparative metabolite profiling was performed on the culture supernatant of the host and deletion strain. As host strain, *P. chrysogenum* DS54555 was used, which is derived from the industrial DS17690 strain, and that lacks the *ku70* gene to make it competent for homologous recombination. The

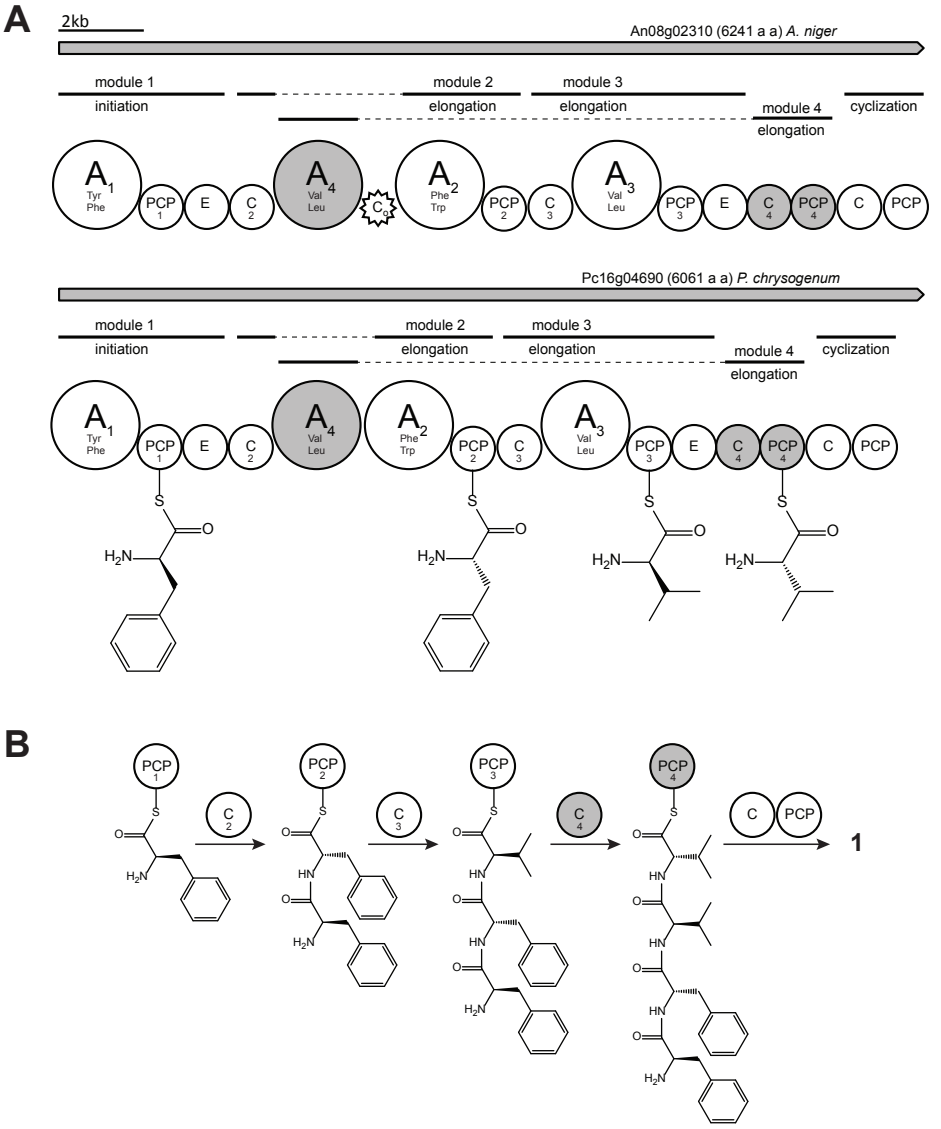


Figure 3. Model for the biosynthesis of compound **1** in *P. chrysogenum* and *A. niger* by the HcpA NRPS.
A: Binding of the monomers on the carrier protein domains. Both isolated condensation and thiolation domains C_4 and T_4, missing a preceding adenylation domain, are assumed to correspond to the adenylation domain A_4, located upstream.
B: Proposed assembly of compound **1** including condensation domains presumed for catalyzing the formation of the peptide bond.

DS54555 strain also lacks the multiple penicillin biosynthetic genes clusters in order to facilitate the detection of unknown secondary metabolites in the culture broth as the profile is no longer dominated by β-lactams. *PcHcpA* was deleted by homologous recombination using the deletion plasmid pDEST R_4-R_3p (Figure S1) containing

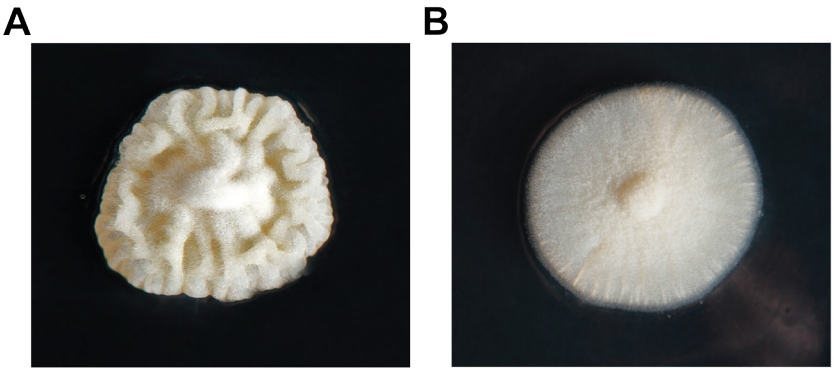


Figure 4. Colonies of P. chrysogenum strain DS54555
A: Colony of host strain showing a wrinkled surface B: Colony of *ΔPcHcpA* strain with a smooth surface

the flanking regions of *PcHcpA* and the phleomycin resistance gene. Colonies were selected on phleomycin containing agar plates where the mutant colonies showed smooth phenotypic characteristics compared to the wrinkled surface of the colonies of the parental strain (Figure 4). The deletion of the *PcHcpA* gene was confirmed by southern blot hybridization (Figure 5A).
The host and *ΔPcHcpA* strain were grown for 168 hours in secondary metabolite production medium (SMP Medium) followed by comparative metabolite analysis using HPLC-UV-MS. Several secondary metabolites were found to be present in the host but absent in the deletion strain (Figure 6, Table S1). These compounds could be classified into two groups according their chemical structure. The first group consists of ten cyclic tetrapeptides, which were identified using HPLC-MS^n, NMR and a synthetic standard (Figure 1). Upon excitation, cyclic tetrapeptides undergo ring opening in the mass spectrometer resulting in four linear tetrapeptides which can be sequenced in a similar fashion as linear peptides. By following the sequential loss of amino acids from b-ions of each generated linear tetrapeptide, the sequence of their cyclic origin could be determined (Figure S2). The identities of the involved amino acids, corresponding to the losses in the mass spectrometer, as well as their sequence were additionally confirmed by ^{1}H-NMR and ^{13}C-NMR experiments for the compounds **1** and ***2*** (Table S2 and S3). To discriminate the amino acid isoleucine from its isomer leucine, which is represented by a loss of 113 Da ($C_6H_{11}ON$) in the mass spectra of compounds **5**-**10**, ^{1}H-, ^{13}C- and various 2D-NMR experiments were conducted (Figure S3, Table S4). Overall, cyclic tetrapeptides obtained from metabolic profiling contain various combinations of the five amino acids valine, isoleucine, phenylalanine, tyrosine and tryptophan making them extremely hydrophobic. They are arranged in a common sequence in which two aromatic amino acids are followed by two aliphatic amino acids. The absolute stereochemistry of compound **2** was confirmed by spiking its synthetic standard to a natural extract which did not lead to additional signals in the ^{1}H-NMR spectra, whereas the intensity of the main signals increased as compared to the impurities (Figure S4A and B). In addition, superimposing the HMBC spectra of the natural and the synthetic sample of peptide **2** illustrated identical correlations and shifts (Figure S5). Furthermore, retention time and MS^2 fragmentation did not show differences between the extracted compounds and their synthesized standards. This leads to the conclusion that not only

the sequence of amino acids is identical in the synthetic and natural peptide, but also the chirality of the individual amino acids. Therefore, the cyclic tetrapeptide **2** has the same stereochemistry as observed for **1** (Studer, 1969) with the first amino acid of an aliphatic and aromatic pair in D- and the second amino acid in L-form.
The second category of identified compounds consists of 18 linear tetrapeptides which are comprised of the same five amino acids as their cyclic analogues (Table S1 and S5). Similar to the cyclic peptides, each linear tetrapeptide contains two aliphatic and two aromatic amino acids in different arrangements yielding various isomeric structures. Due to similar chromatographic properties and an identical mass-over-charge ratio, these isomers are represented as a group in data obtained from metabolic profiling (Figure S6A and B). Their structure elucidation is challenging as

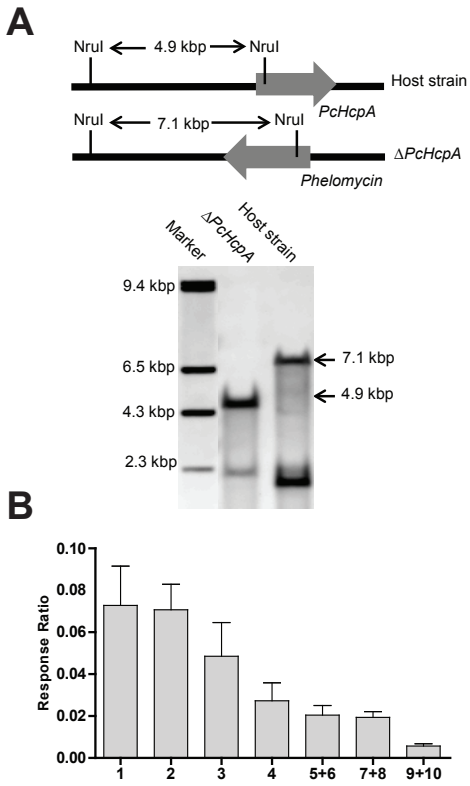


Figure 5. Southern blot analysis for *PcHcpA* deletion and concentration of cyclic tetrapeptides in culture broth of host strain
A: Southern blot hybridization validating the complete gene deletion of *PcHcpA*.
B: Internal standard corrected concentrations of the cyclic tetrapeptides **1**-**10** present in the culture broth of the host strain of *P. chrysogenum* grown for 168 hours. Isomers are presented together as no chromatographic separation was obtained during profiling. No cyclic tetrapeptides could be found in the deletion strain.

minor fragments can be attributed to a fragmentation of low abundant linear tetrapeptides as well as to possible sequence scrambling of major linear tetrapeptides which was reported for similar linear peptides (Bleiholder et al., 2008). To separate and sequence these isomeric tetrapeptides and to prevent possible sequence scrambling, their N-terminus was derivatized using 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate (AQC) (Figure S6C and D) (Cohen and Michaud, 1993). *De novo*

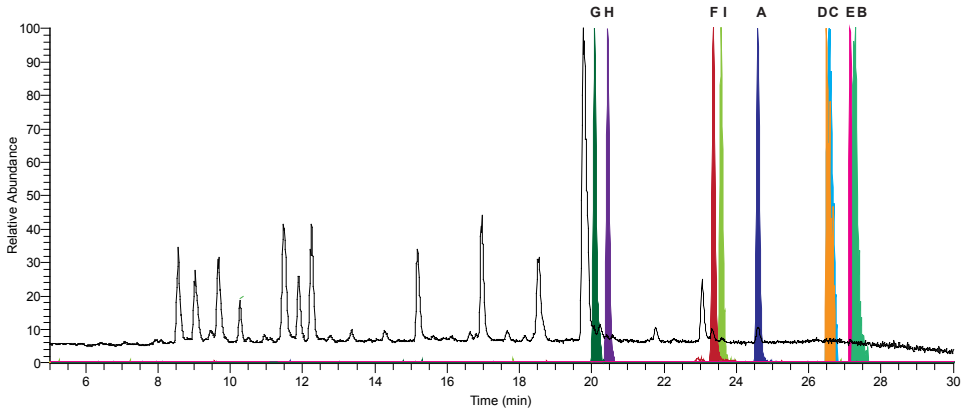


Figure 6. HPLC-MS elution profiles of the internal standard reserpine and the highest abundant identified peptides
Total ion chromatogram (TIC, black) and normalized extracted ion chromatograms (EIC, colored) of the protonated molecule ($[M+H]^+$ with 5 ppm accuracy) of the used internal standard reserpine (A), the four highest abundant cyclic tetrapeptides **1** (B), **2** (C), **3** (D), **4** (E) and linear tetrapeptides **11**-**13** (F), **14**-**16** (G), **17**-**19** (H), **20**-**22** (I) present in the culture broth of *P. chrysogenum*.

peptide sequencing was performed by following the consecutive amino acid losses of various b-ions using multiple-stage fragmentation mass spectrometry (Figure S6E and Table S5). Similar to the cyclic peptides, each linear tetrapeptide incorporates two consecutive aromatic and/or two consecutive aliphatic amino acids.
In conclusion, several linear and cyclic tetrapeptides with similar structural features were found to be present in the host but absent in the deletion strain which shows that they originate from one single NRPS, namely PcHcpA.

Expression of the PcHcpA gene and secondary metabolite production
To examine the expression of the *PcHcpA* gene, the host strain was grown for up to 216 hours in secondary SMP medium. Samples were collected for total mRNA extraction and extracellular metabolites analysis during growth. Metabolite concentrations were determined by HPLC-UV-MS, while transcript levels were determined by quantitative PCR using γ-actin as a reference gene. A high expression level of the PcHcpA gene was observed between 120 and 168 hours of growth, which was paralleled by a 12 times increase in the concentration of cyclic tetrapeptide **1** (Figure 2A and B) in the medium. In general, the concentration of cyclic tetrapeptides was exceptionally high around 168 hours of growth, while the concentration of the linear tetrapeptides increased with time. These data suggest that the linear tetrapeptides are derived from the cyclic tetrapeptides by degradation.

Discussion

Here we have demonstrated that the NRPS PcHcpA is responsible for the biosynthesis of cyclic hydrophobic tetrapeptides. Bioinformatics analysis of PcHcpA shows an unusual domain architecture in comparison to classical linear NRPSs with no products associated so far. However, through the deletion of the *PcHcpA* gene and comparative metabolite profiling, ten cyclic tetrapeptides were identified, including

the previously described secondary metabolite fungisporin **1** (Studer, 1969). According to the non-ribosomal code in combination with discovered cyclic products, the first NRPS module is specific for phenylalanine. Due to the adjacent E domain, responsible for epimerization of the activated amino acid, a D-configuration is expected as observed for product **1** (Figure 1). The next two modules of the NRPS contain an unusual architecture in which PCP_2 and C_3 are flanked by two neighboring adenylation domains A_2 and A_4. As adenylation domain A_4 shows high specificity towards valine and adenylation domain A_2 shows high homology towards A_1, both adenylation domains activate phenylalanine as supported by the structure of **1**. Due to the lack of an adjacent PCP domain for A_4 and a missing C domain for A_2, which are necessary for loading and condensation of the substrate, no module activity is expected according the classical C-A-PCP geometry (Figure 3A). Surprisingly, both products **1** and **2** show the incorporation of L-phenylalanine at the second position. As module 2 is predicted to be the only module capable to catalyze the incorporation of L-phenylalanine, A_2 indeed must be active. Consequently, it seems likely that A_4 is skipped, leaving the N-terminal C_2 domain of module 2 to catalyze the condensation of the first amino acid from module 1 to the second amino acid from module 2, as observed for the products **1** and **2**. The third module of PcHcpA contains the domains C_3, A_3, PCP_3 and E arranged in a common linear order. The A_3 domain is predicted to activate valine which agrees with the peptide sequence of **1** and **2**, as D-valine is their third amino acid. The fourth module of the NRPS is an incomplete module consisting of C_4 and PCP_4. Due to a missing preceding A domain no activity is expected. However, the chemical structures of both cyclic tetrapeptides **1** and **2** show the incorporation of L-valine as fourth amino acid in their peptide sequence. As A_4 is the only domain predicted to be specific to valine without an adjacent epimerization domain, it is very likely that this domain is a 'trans-acting' A domain that interacts with C_4 and PCP_4 to add the last amino acid to the tetrapeptide. A similar architectural flexibility has been observed in the biosynthesis of yersiniabactin, in which one A domain, located in HMWP2, loads three PCPs located on different modules (Gehring et al., 1998a; Gehring et al., 1998b; Suo et al., 2001). As the linear domain organization of PcHcpA does not reflect a linear assembly of substrate incorporation into the final product, non-linear interactions are deduced. Although A_4, C_4 and PCP_4 are not in a consecutive sequence on a genomic level, they might still be closely arranged in the final three-dimensional enzymatic structure. Structural characterization would be necessary to determine spatial proximity. Finally, after the incorporation of L-valine into the peptide chain, the C and PCP domain of the last module catalyze the cyclization of the peptide leading to the final cyclic structure, as previously observed in other NRPS systems (Gao et al., 2012; Keating et al., 2001).

Next to the production of **1** and **2**, eight additional lower abundant cyclic tetrapeptides were identified to be present in the host strain and absent in the deletion strain (Figure 1). They show a similar peptide sequence as **1**, containing two aromatic amino acids followed by two aliphatic amino acids. Although stereochemical information is only available for the compounds **1** and **2**, it can be assumed that each of the cyclic tetrapeptides contains an aliphatic and aromatic amino acid in the D configuration, more specifically at the first and third position. These assumptions in combination with the stereochemical structure of **1** and **2** lead to the conclusion,

that each adenylation domain of PcHcpA shows specificity towards more than one precursor amino acid with A_1 being specific towards phenylalanine and tyrosine and A_2 being specific towards phenylalanine and to a lesser extent to tryptophan (Table S6). Together with the two aliphatic amino acid selecting adenylation domains A_3 and A_4, which preferably activate valine before isoleucine, 16 cyclic tetrapeptide combinations are theoretically possible. However, only ten of these were detected in the fermentation broth of *P. chrysogenum* confirming a different degree of specificity towards their precursors. Based on a similar chemical scaffold of identified compounds **1**-**10** to the tetrapeptides *cyclo*-(*N*-MePhe-Ile)$_2$, *cyclo*-(*N*-MePhe-Val)$_2$ and *cyclo*-(*N*-MePhe-val-*N*-MePhe-Ile) reported from *Onychocola sclerotic*, cardiac channel blocking activities can be expected for the hydrophobic cyclic peptides presented here (Perez-Victoria et al., 2012). In addition, the colonies of the *ΔPcHcpA* strain lost the ability to produce a wrinkled surface leading to a rather smooth appearance (Figure 4). As this change is attributed to the deletion of the *PcHcpA* gene, the hydrophobic cyclic peptides **1**-**10** need to be involved. Possibly, these molecules function analogous to hydrophobins in altering the surface properties and influencing aerial growth. The exact function is, however, still unclear.

Next to cyclic tetrapeptides several highly abundant linear tetrapeptides could be observed in the cultural broth of the host strain that were absent in the deletion strain. To each of the cyclic tetrapeptides, several linear tetrapeptides with the same sequence were present. For instance, in addition to the cyclic tetrapeptide **1** with the sequence *cyclo*-(Phe-Phe-Val-Val), three linear tetrapeptides with the sequences Phe-Val-Val-Phe, Val-Phe-Phe-Val and Phe-Phe-Val-Val could be found at different ratios. Their concentration increased over time in the media while their cyclic counterpart decreased after 168 hours (Figure 2B). This leads to the conclusion that the linear peptides originate from the degradation of their cyclic counterparts by hydrolysis of their peptide bonds, which was observed exclusively between two aromatic, two aliphatic or an aliphatic followed by an aromatic amino acid (Table S1 and S5). Linear tetrapeptides with a N-terminal aliphatic amino acid and a C-terminal aromatic amino acid were not detected, leading to the conclusion that cleavage of this bond is not favorable. As cyclic tetrapeptides are relatively stable towards chemical and thermal degradation, enzymatic hydrolysis might be most probable.

PcHcpA shows 54% amino acid sequence identity with its orthologous AnHcpA from *A. niger* with exactly the same module organization (Figure 3A). Furthermore, all cyclic products **1**-**10** present in *P. chrysogenum* could also be found in the supernatant of an *A. niger* strain while they were absent in the *AnHcpA* deletion strain (unpublished data). Therefore, it is concluded that AnHcpA is involved in production of all cyclic metabolites (**1**-**10**) in *A. niger*. A small difference exists in the organization of the PcHcpA and AnHcpA proteins with a short additional amino acid sequence present between domains A_4 and A_2 in AnHcpA. This sequence showed limited homology to a condensation domain and appears as an incomplete condensation domain. Hence, one may deduce that this noncanonical situation might have evolved quite recently. Perhaps, a complete C-A-PCP-C-A-PCP module structure was present before, but degenerated after new interactions between the domains evolved.

References

Ali, H., Ries, M.I., Nijland, J.G., Lankhorst, P.P., Hankemeier, T., Bovenberg, R.A.L., Vreeken, R.J., and Driessen, A.J.M. (2013). A branched biosynthetic pathway is involved in production of roquefortine and related compounds in Penicillium chrysogenum. PLOS ONE 8, e65328.

Alvarez, E., Cantoral, J.M., Barredo, J.L., Diez, B., and Martin, J.F. (1987). Purification to homogeneity and characterization of acyl coenzyme A:6-aminopenicillanic acid acyltransferase of Penicillium chrysogenum. Antimicrob. Agents Chemother. 31, 1675-1682.

Baltz, R.H. (2011). Function of MbtH homologs in nonribosomal peptide biosynthesis and applications in secondary metabolite discovery. J. Ind. Microbiol. Biotechnol. 38, 1747-1760.

Bleiholder, C., Osburn, S., Williams, T.D., Suhai, S., Van Stipdonk, M., Harrison, A.G., and Paizs, B. (2008). Sequence-scrambling fragmentation pathways of protonated peptides. J. Am. Chem. Soc. 130, 17774-17789.

Cohen, S.A., and Michaud, D.P. (1993). Synthesis of a fluorescent derivatizing reagent, 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate, and its application for the analysis of hydrolysate amino acids via high-performance liquid chromatography. Anal. Biochem. 211, 279-287.

Conti, E., Stachelhaus, T., Marahiel, M.A., and Brick, P. (1997). Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. EMBO J. 16, 4174-4183.

Doekel, S., Coeffet-Le Gal, M.F., Gu, J.Q., Chu, M., Baltz, R.H., and Brian, P. (2008). Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. Microbiology 154, 2872-2880.

Fischbach, M.A., and Walsh, C.T. (2006). Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. 106, 3468-3496.

Frisvad, J.C., Smedsgaard, J., Larsen, T.O., and Samson, R.A. (2004). Mycotoxins, drugs and other extrolites produced by species in Penicillium subgenus Penicillium. Studies in Mycology 49, 201-241.

Gao, X., Haynes, S.W., Ames, B.D., Wang, P., Vien, L.P., Walsh, C.T., and Tang, Y. (2012). Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. Nat. Chem. Biol. 8, 823-830.

Garcia-Estrada, C., Ullan, R.V., Albillos, S.M., Fernandez-Bodega, M.A., Durek, P., von Dohren, H., and Martin, J.F. (2011). A single cluster of coregulated genes encodes the biosynthesis of the mycotoxins roquefortine C and meleagrin in Penicillium chrysogenum. Chem. Biol. 18, 1499-1512.

Gehring, A.M., DeMoll, E., Fetherston, J.D., Mori, I., Mayhew, G.F., Blattner, F.R., Walsh, C.T., and Perry, R.D. (1998a). Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by Yersinia pestis. Chem. Biol. 5, 573-586.

Gehring, A.M., Mori, I., Perry, R.D., and Walsh, C.T. (1998b). The nonribosomal peptide synthetase HMWP2 forms a thiazoline ring during biogenesis of yersiniabactin, an iron-chelating virulence factor of Yersinia pestis. Biochemistry 37, 11637-11650.

Grunewald, J., and Marahiel, M.A. (2006). Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. Microbiol. Mol. Biol. Rev. 70, 121-146.

Harju, S., Fedosyuk, H., and Peterson, K.R. (2004). Rapid isolation of yeast genomic DNA: Bust n' Grab. BMC Biotechnol. 4, 8.

Keating, T.A., Ehmann, D.E., Kohli, R.M., Marshall, C.G., Trauger, J.W., and Walsh, C.T. (2001). Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-S-enzyme breakdown in antibiotic and siderophore biosynthesis. Chembiochem 2, 99-107.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47, 736-741.

Kolar, M., Punt, P.J., van den Hondel, C.A., and Schwab, H. (1988). Transformation of Penicillium chrysogenum using dominant selection markers and expression of an Escherichia coli lacZ fusion gene. Gene 62, 127-134.

Kovalchuk, A., Weber, S.S., Nijland, J.G., Bovenberg, R.A., and Driessen, A.J. (2012). Fungal ABC transporter deletion and localization analysis. Methods Mol. Biol. 835, 1-16.

Mootz, H.D., Schwarzer, D., and Marahiel, M.A. (2002). Ways of assembling complex natural products on modular nonribosomal peptide synthetases. Chembiochem 3, 490-504.

Nijland, J.G., Kovalchuk, A., van den Berg, M.A., Bovenberg, R.A., and Driessen, A.J. (2008). Expression of the transporter encoded by the cefT gene of Acremonium chrysogenum increases cephalosporin production in Penicillium chrysogenum. Fungal Genet. Biol. 45, 1415-1421.

Noga, M.J., Dane, A., Shi, S., Attali, A., van Aken, H., Suidgeest, E., Tuinstra, T., Muilwijk, B., Coulier, L., Luider, T. et al. (2012). Metabolomics of cerebrospinal fluid reveals changes in the central nervous system metabolism in a rat model of multiple sclerosis. Metabolomics 8, 253-263.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., Schaap, P.J., Turner, G., de Vries, R.P., Albang, R., Albermann, K. et al. (2007). Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. Nat. Biotechnol. 25, 221-231.

Perez-Victoria, I., Martin, J., Gonzalez-Menendez, V., de Pedro, N., El Aouad, N., Ortiz-Lopez, F.J., Tormo, J.R., Platas, G., Vicente, F., Bills, G.F. et al. (2012). Isolation and structural elucidation of cyclic tetrapeptides from Onychocola sclerotica. J. Nat. Prod. 75, 1210-1214.

Rottig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 39, W362-7.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press. 2nd ed.

Schwarzer, D., Finking, R., and Marahiel, M.A. (2003). Nonribosomal peptides: from genes to products. Nat. Prod. Rep. 20, 275-287.

Shaw-Reid, C.A., Kelleher, N.L., Losey, H.C., Gehring, A.M., Berg, C., and Walsh, C.T. (1999). Assembly line enzymology by multimodular nonribosomal peptide synthetases: the thioesterase domain of E. coli EntF catalyzes both elongation and cyclolactonization. Chem. Biol. 6, 385-400.

Stevens, B.W., Lilien, R.H., Georgiev, I., Donald, B.R., and Anderson, A.C. (2006). Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. Biochemistry 45, 15495-15504.

Studer, R.O. (1969). Synthesis and structure of fungisporin. Experientia 25, 899.

Suo, Z., Tseng, C.C., and Walsh, C.T. (2001). Purification, priming, and catalytic acylation of carrier protein domains in the polyketide synthase and nonribosomal peptidyl synthetase modules of the HMWP1 subunit of yersiniabactin synthetase. Proc. Natl. Acad. Sci. U. S. A. 98, 99-104.

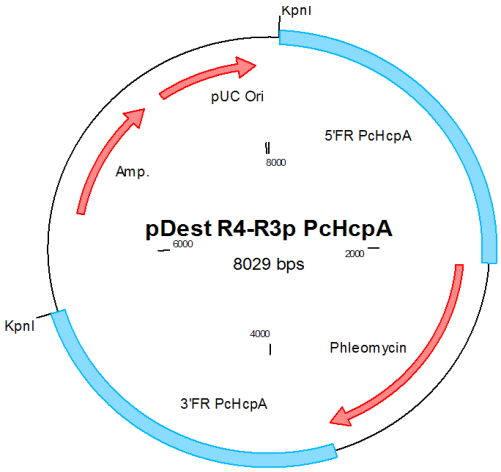
van den Berg, M.A., Albang, R., Albermann, K., Badger, J.H., Daran, J.M., Driessen, A.J., Garcia-Estrada, C., Fedorova, N.D., Harris, D.M., Heijne, W.H. et al. (2008). Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. Nat. Biotechnol. 26, 1161-1168.

van Peij, N.N.M.E., Hans M., Beishuizen M., Schipper D., van der Hoeven R.A.M., Schouten, O.L. (2012). A method for the production of a compound of interest. WO2012001169

Wyatt, M.A., Mok, M.C., Junop, M., and Magarvey, N.A. (2012). Heterologous expression and structural characterisation of a pyrazinone natural product assembly line. Chembiochem 13, 2408-2415.

Zhang, K., Nelson, K.M., Bhuripanyo, K., Grimes, K.D., Zhao, B., Aldrich, C.C., and Yin, J. (2013). Engineering the Substrate Specificity of the DhbE Adenylation Domain by Yeast Cell Surface Display. Chem. Biol. 20, 92-101.

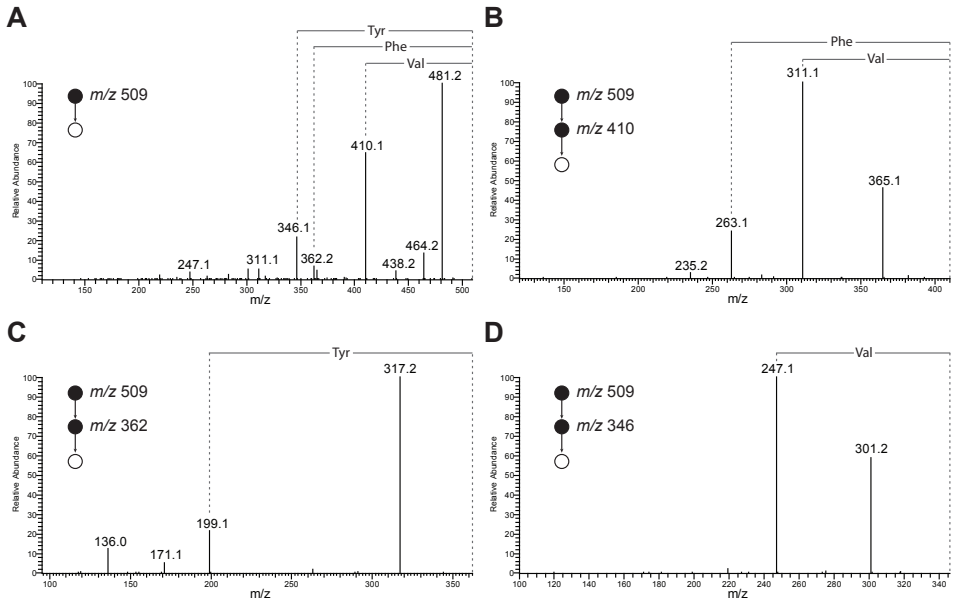
Supplemental Information



Supplementary Figure 1.
Deletion plasmid used for deleting *HcpA* in *P. chrysogenum*

peptide	peptide sequence	RT	formula	acquired mass $[M+H]^+$	accuracy
		min		m/z	ppm
1	*cyclo*-(FFVV)	27.33	C28H36N4O4	493.2806	-0.673
2	*cyclo*-(YFVV)	26.63	C28H36N4O5	509.2757	-0.288
3	*cyclo*-(YWVV)	26.59	C30H37N5O5	548.2868	0.099
4	*cyclo*-(FWVV)	27.21	C30H37N5O4	532.2916	-0.434
5	*cyclo*-(FFVI)	27.56	C29H38N4O4	507.2962	-0.753
6	*cyclo*-(FFIV)	27.56	C29H38N4O4	507.2962	-0.753
7	*cyclo*-(YWVI)	26.86	C31H39N5O5	562.3017	-1.237
8	*cyclo*-(YWIV)	26.86	C31H39N5O5	562.3017	-1.237
9	*cyclo*-(YFVI)	26.90	C29H38N4O5	523.2914	-0.185
10	*cyclo*-(YFIV)	26.90	C29H38N4O5	523.2914	-0.185
11	FVVF	23.34	C28H38N4O5	511.2912	-0.580
12	VFFV	23.34	C28H38N4O5	511.2912	-0.580
13	FFVV	23.34	C28H38N4O5	511.2912	-0.580
14	YFVV	20.06	C28H38N4O6	527.2864	-0.022
15	VYFV	20.06	C28H38N4O6	527.2864	-0.022
16	FVVY	20.06	C28H38N4O6	527.2864	-0.022
17	YWVV	20.40	C30H39N5O6	566.2972	-0.195
18	VYWV	20.40	C30H39N5O6	566.2972	-0.195
19	WVVY	20.40	C30H39N5O6	566.2972	-0.195
20	VFWV	23.57	C30H39N5O5	550.3023	-0.174
21	FWVV	23.57	C30H39N5O5	550.3023	-0.174
22	WVVF	23.57	C30H39N5O5	550.3023	-0.174
23	FVIF	24.66	C29H40N4O5	525.3077	1.053
24	FIVF	24.66	C29H40N4O5	525.3077	1.053
25	FVIY	21.43	C29H40O6N4	541.3023	0.441
26	IYFV	21.43	C29H40O6N4	541.3023	0.441
27	FIVY	21.43	C29H40O6N4	541.3023	0.441
28	VYFI	21.43	C29H40O6N4	541.3023	0.441

Supplementary Table 1.
Retention time, formula and acquired *m/z* of cyclic and linear tetrapeptides obtained from metabolic profiling. Isomers have the same retention time as they could not be chromatographically separated during profiling.



Supplementary Figure 2. Multiple-stage fragmentation for *de novo* sequencing of compound **2** based on sequential amino acid losses

A: MS^2 fragmentation spectra of the cyclic tetrapeptide **2** *cyclo*-(*d*-Tyr-*l*-Phe-*d*-Val-*l*-Val). Due to ring opening of the cyclic peptide in the mass spectrometer at different positions, three different amino acid losses occurred, yielding different b_3-ions.

B-D: MS^3 fragmentation spectra obtained by further fragmenting b_3-ions from MS^2 showing b_2-ions used for peptide sequencing.

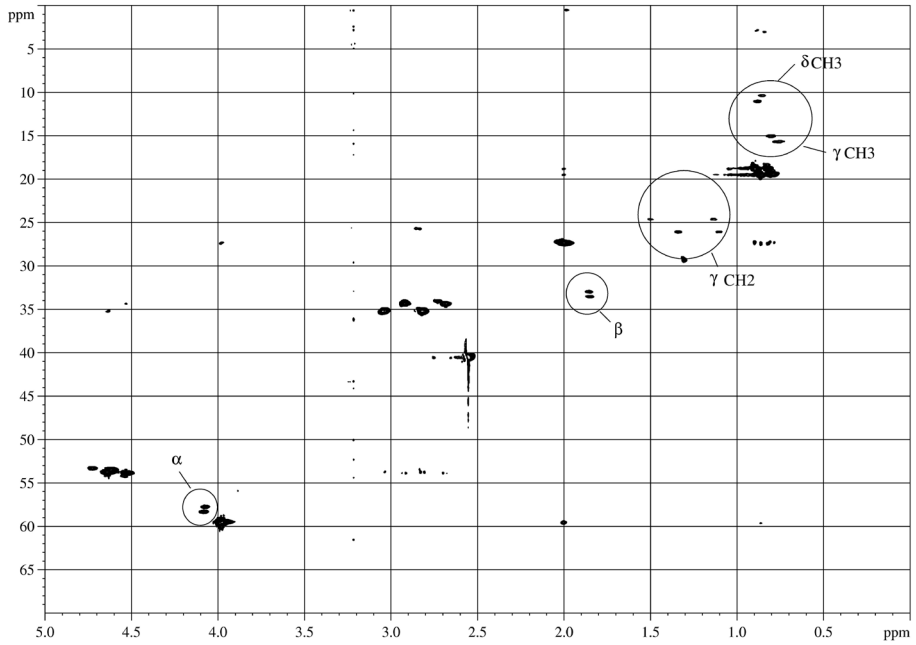
The cyclic tetrapeptides **1** and **3**-**10** were identified accordingly.

amino acid	position	δ (^{1}H)	Position	δ(^{13}C)
d-Tyr	α	4.53	α	54.10
	β1	2.93	β	34.59
	β2	2.69	γ	128.43
	δ	6.99	δ	130.32
	ε	6.65	ε	115.63
	NH	7.82	ζ	156.35
	OH	8.99	C=O	173.35
l-Phe	α	4.62	α	53.97
	β1	3.05	β	35.47
	β2	2.80	γ	138.48
	δ	7.20	δ	129.41
	ε	7.21	ε	128.68
	ζ	7.26	ζ	126.72
	NH	7.80	C=O	173.56
d-Val	α	4.00	α	59.90
	β	2.00	β	27.66
	γ1	0.87	γ1	19.76
	γ2	0.82	γ2	19.01
	NH	7.58	C=O	173.81
l-Val	α	3.98	α	59.71
	β	2.00	β	27.51
	γ1	0.90	γ1	19.73
	γ2	0.80	γ2	19.07
	NH	7.62	C=O	173.32

Supplementary Table 2.
^{1}H and ^{13}C-NMR chemical shifts of synthetically produced compound **2** with sequence *cyclo*-(*d*-Tyr-*l*-Phe-*d*-Val-*l*-Val) in DMSO at 340 K. As synthetically and naturally produced **2** show identical NMR spectra, only chemical shifts for the synthetically produced compound are shown. δ_{DMSO} (^{1}H/^{13}C) = (2.55/40.50), (δ in ppm).

amino acid	position	δ (^{1}H)	Position	δ(^{13}C)
d-Phe	α	4.62*	α	53.69
	β1	3.05*	β	35.29
	β2	2.80*	γ	138.40
	δ	7.20*	δ	129.43
	ε	7.21*	ε	128.66
	ζ	7.26*	ζ	126.74
	NH	7.88	C=O	173.16
l-Phe	A	4.62*	α	54.03
	β1	3.05*	β	35.56
	β2	2.80*	γ	138.44
	δ	7.20*	δ	129.42
	ε	7.21*	ε	128.68*
	ζ	7.26*	ζ	126.72*
	NH	7.78	C=O	173.56*
d-Val	α	4.00*	α	59.88
	β	2.00*	β	27.61
	γ1	0.86	γ1	19.77
	γ2	0.81	γ2	19.01*
	NH	7.58	C=O	173.79
l-Val	α	3.98*	α	59.73
	β	2.00*	β	27.54
	γ1	0.89	γ1	19.67
	γ2	0.78	γ2	19.06
	NH	7.58	C=O	173.36

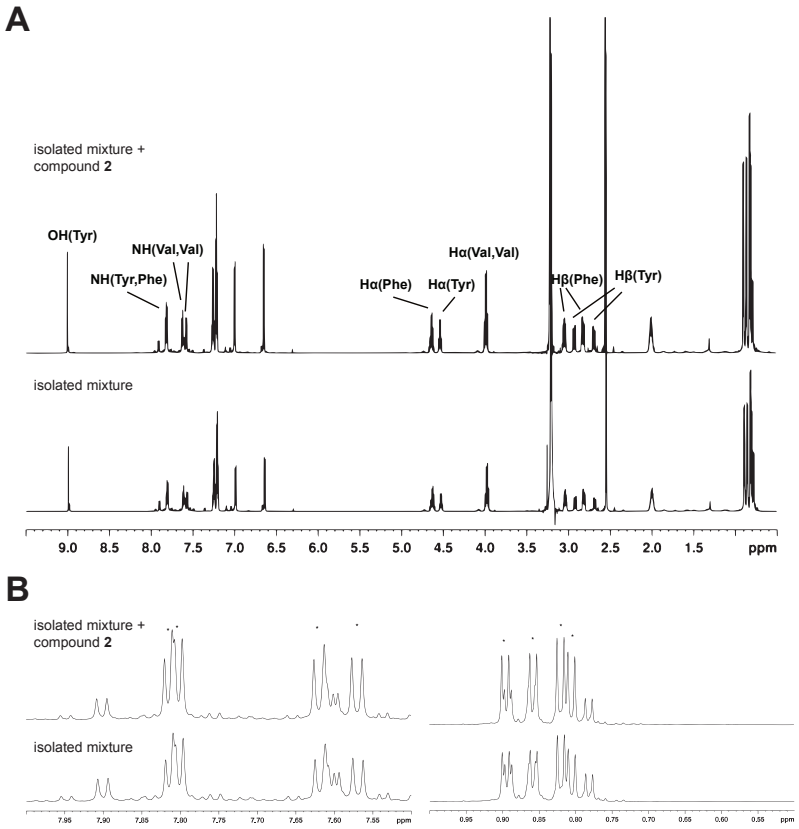
Supplementary Table 3.
^{1}H and ^{13}C chemical shifts of naturally produced compound **1** with sequence *cyclo*-(*d*-Phe-*l*-Phe-*d*-Val-*l*-Val) present in a mix of various cyclic tetrapeptides in DMSO acquired at 340 K. Signals overlapping with the highest abundant compound **2** are indicated (*). δDMSO (^{1}H/^{13}C) = (2.55/40.50), (δ in ppm).



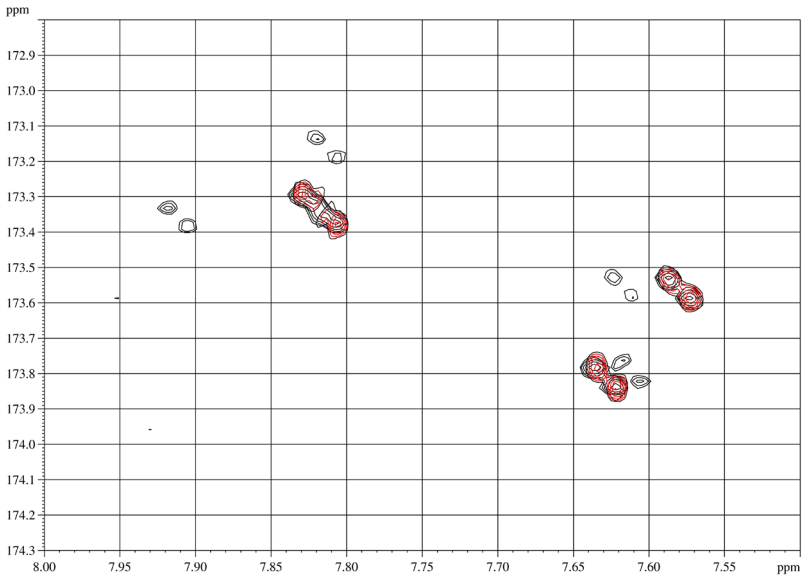
Supplementary Figure 3.
HSQC spectrum of isolated mix of cyclic tetrapeptides in DMSO used for the identification of isoleucine present in the minor abundant products **5** and **9**. Signals corresponding to isoleucine are indicated with circles. Conducted TOCSY, COSY and HMBC experiments further confirm this conclusion (data not shown). ^{1}H and ^{13}C chemical shifts are shown in Supplementary Table 4.

amino acid	position	δ (^{1}H)	Position	δ(^{13}C)
Ile in **5**	α	4.1	α	57.8
	β	1.85	β	34.0
	γ1	0.76	γ1	15.0
	γ2	1.3	γ2	26.5
	γ2´	1.1		
	δ	0.9	δ	11.0
	NH	n.o.	C=O	n.o.
Ile in **9**	α	4.1	α	57.5
	β	1.85	β	33.5
	γ1	0.74	γ1	16.0
	γ2	1.5	γ2	24.5
	γ2´	1.15		
	δ	0.85	δ	10.5
	NH	n.o.	C=O	n.o.

Supplementary Table 4.
^{1}H and ^{13}C chemical shifts of isoleucine in compound **5** with sequence *cyclo*-(Phe-Phe-Val-Ile) and compound **9** with sequence *cyclo*-(Phe-Phe-Val-Ile) present in a extracted mix of cyclic tetrapeptides. NMR signals corresponding to C=O and NH as well as remaining amino acid signals are not observed (n.o.) due to overlap with the main constituents. δ_{DMSO} (^{1}H/^{13}C) = (2.55/40.50).



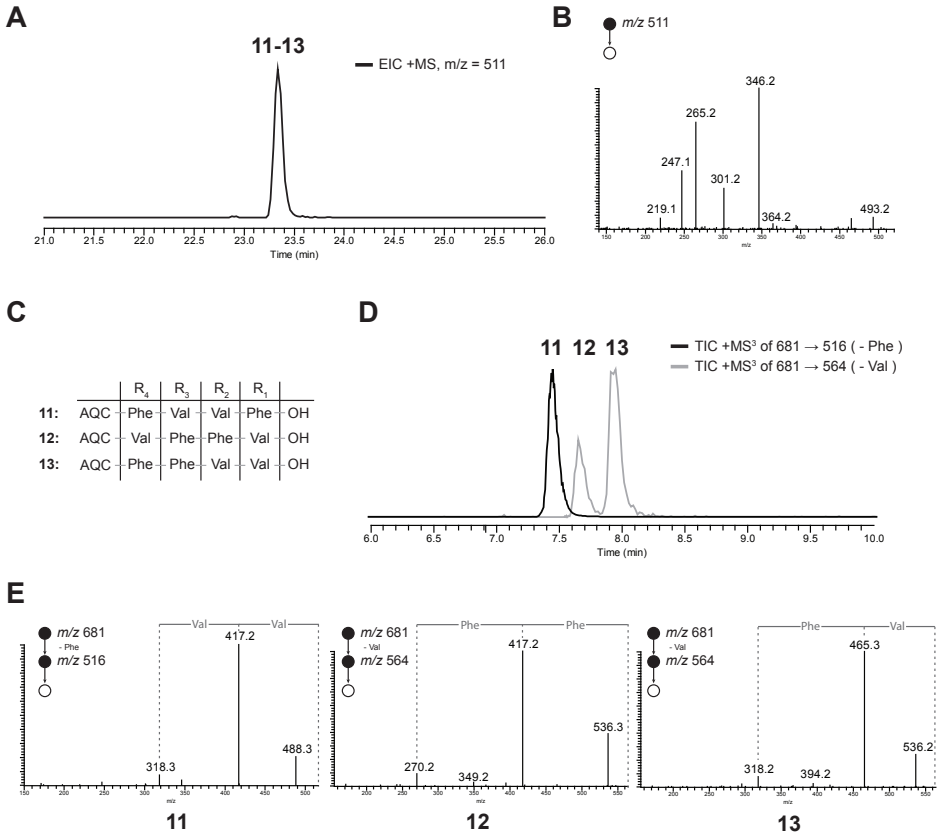
Supplementary Figure 4.
A: ^{1}H-NMR spectrum of a precipitate of various cyclic tetrapeptides containing primarily peptide **1** and **2** (bottom). Synthetic compound **2** spiked to the precipitated mix of various cyclic tetrapeptides in DMSO at 340 K (top). Signals corresponding to **2** were increased as compared to the impurities whereas additional signals did not appear.
B: Zoomed regions of ^{1}H-NMR spectrum of natural precipitate (bottom) and precipitate spiked with compound **2** (top). Signals which increased after spiking are indicated (*).



Supplementary Figure 5.
Superimposed HMBC spectra of synthetic compound **2** (red) and an isolated mix of cyclic tetrapeptides containing naturally produced compound **2** (black). Correlations between NH and CO are shown which indicate identical shifts for both samples. Assignments can be found in Supplementary Table 2.

linear peptide	underivatized sequence	RT	formula derivatized peptide	$[M+H]^+$	b_3 ion in MS^2	b_2 ion in MS^3 derived from b_3 ion	b_1 ion in MS^3 derived from b_3 ion
		min		m/z	m/z (loss)	m/z (loss)	m/z (loss)
11	FVVF	7.50	C38H44N6O6	681	516 (-F)	417 (-V)	318 (-VV)
12	VFFV	7.71	C38H44N6O6	681	564 (-V)	417 (-F)	270 (-FF)
13	FFVV	7.98	C38H44N6O6	681	564 (-V)	465 (-V)	318 (-FV)
14	YFVV	4.46	C38H44N6O7	697	580 (-V)	481 (-V)	334 (-FV)
15	VYFV	3.28	C38H44N6O7	697	580 (-V)	433 (-F)	270 (-YF)
16	FVVY	3.85	C38H44N6O7	697	516 (-Y)	417 (-V)	318 (-VV)
17	YWVV	4.20	C40H45N7O7	736	619 (-V)	520 (-V)	334 (-WV)
18	VYWV	2.55	C40H45N7O7	736	619 (-V)	433 (-W)	270 (-YW)
19	WVVY	3.64	C40H45N7O7	736	555 (-Y)	456 (-V)	357 (-VV)
20	VFWV	7.01	C40H45N7O6	720	603 (-V)	417 (-W)	270 (-FW)
21	FWVV	7.78	C40H45N7O6	720	603 (-V)	504 (-V)	318 (-WV)
22	WVVF	7.41	C40H45N7O6	720	555 (-F)	546 (-V)	357 (-VV)
23	FVIF	8.24	C39H46N6O6	695	530 (-F)	417 (-I)	318 (-VI)
24	FIVF	8.38	C39H46N6O6	695	530 (-F)	431 (-V)	318 (-IV)
25	FVIY	5.58	C39H46N6O7	711	530 (-Y)	417 (-I)	318 (-VI)
26	IYFV	4.62	C39H46N6O7	711	594 (-V)	447 (-F)	284 (-YF)
27	FIVY	6.14	C39H46N6O7	711	530 (-Y)	431 (-V)	318 (-IV)
28	VYFI	5.52	C39H46N6O7	711	580 (-I)	433 (-F)	270 (-YF)

Supplementary Table 5.
Sequencing data leading to the identification of linear tetrapeptides after AQC derivatization. Multiple-stage fragmentation was used for determination of b-ions for *de-novo* peptide sequencing.



Supplementary Figure 6. Sequencing of the linear isomers **11** (Phe-Phe-Val-Val), **12** (Val-Phe-Phe-Val) and **13** (Phe-Val-Val-Phe).

A: Extracted ion chromatogram (EIC) of the linear tetrapeptides **11**, **12** and **13** using the profiling method. No chromatographic separation could be achieved.

B: MS^2 fragmentation spectrum of unseparated linear tetrapeptides **11**-**13**. Although the spectrum is dominated by fragments originating from **11**, several lower abundant fragments can be found originating from a fragmentation of **12** and **13** or possible sequence scrambling of **11**.

C: Sequence of linear tetrapeptides after N-terminal AQC derivatization to achieve better chromatographic separation and to prevent sequence scrambling.

D: Normalized total ion chromatogram (TIC) of AQC derivatized linear peptides **11**-**13** after first C-terminal amino acid loss showing chromatographic separation and allowing peptide sequencing.

E: Individual MS^3 fragmentation spectra of chromatographically separated derivatized peptides **11**, **12** and **13** showing b_2 and b_1 ions used for peptide sequencing.

The linear peptides **14** – **28** were identified accordingly.

cyclic peptide	specificity of adenylation domain				corresponding linear peptide
	A_1	A_2	A_3	A_4	
1	Phe	Phe	Val	Val	11-13
2	Tyr	Phe	Val	Val	14-16
3	Tyr	Trp	Val	Val	17-19
4	Phe	Trp	Val	Val	20-22
5	Phe	Phe	Val	Ile	23
6	Phe	Phe	Ile	Val	24
7	Tyr	Trp	Val	Ile	*not observed*
8	Tyr	Trp	Ile	Val	*not observed*
9	Tyr	Phe	Val	Ile	25-26
10	Tyr	Phe	Ile	Val	27-28
proposed	Phe	Trp	Val	Ile	*not observed*
proposed	Phe	Trp	Ile	Val	*not observed*
proposed	Phe	Phe	Ile	Ile	*not observed*
proposed	Phe	Trp	Ile	Ile	*not observed*
proposed	Tyr	Phe	Ile	Ile	*not observed*
proposed	Tyr	Trp	Ile	Ile	*not observed*
specificity	Phe/Tyr	Phe/Trp	Val/Ile	Val/Ile	

Supplementary Table 6.
Identified and proposed cyclic and linear tetrapeptides in respect to the adenylation domain specificity in the NRPS PcHcpA.

Chapter 5

Chemoinformatics supported MS^n Comparison Pipeline (CMCP): Towards automated *de novo* structure elucidation using multiple-stage fragmentation tree comparison

Based on

Marco I. Ries, Jeroen Kazius, Hazrat Ali, Arnold J.M. Driessen, Thomas Hankemeier, Theo Reijmers, Rob J. Vreeken
Chemoinformatics supported MS^n Comparison Pipeline (CMCP): Towards automated de novo structure elucidation using multiple-stage fragmentation tree comparison
In preparation for publication

Abstract

Here we present a novel multiple stage fragmentation tree based structure elucidation pipeline for the *de novo* structure identification of small molecules, coined CMCP (Chemoinformatics supported MS^n Comparison Pipeline) which combines sample extraction, fragmentation tree acquisition, database screening and data interpretation. Based on high-resolution multiple-stage fragmentation mass spectra (MS^n), fragment and neutral-loss trees are generated and compared to a MS^n database to extract compounds with similar fragmentation behavior. By identifying their shared fragmentation mechanisms, chemical structures of unknowns can be deduced without them needing to be present in the database. As the structure elucidation presented here is based on structural identification of similar fragments and neutral-losses rather than solely their comparison, confident identification of unknown molecular structures is possible beyond classifications into compound classes. Most important conceptual outcomes of a database query using fragment and the newly introduced neutral-loss tree comparison are discussed showing how this information can ultimately be used to deduce complete chemical (sub)structures of unknown compounds. For their demonstration, the *de novo* identification of the complex metabolites roquefortine C, dehydrohistidyltryptophanyldiketopiperazine (DHTD), roquefortine F and roquefortine D extracted from cultures of *Penicillium chrysogenum* is described in detail.

Introduction

With the prevalent application of metabolomics, several methods and techniques for the simultaneous analysis of a large number of metabolites have been developed which provide detailed information about the metabolome of complex biological samples. However, their structural identification, necessary to describe and interpret cellular processes, is still a major bottleneck in metabolomics (Kind and Fiehn, 2006).
Mass spectrometry (MS) in combination with chromatographic separation techniques is widely used for the analysis of metabolites due to its high sensitivity, low demand on sample purity and rapid analysis time (Villas-Boas, et al., 2005). Especially liquid chromatography MS (LC-MS), which is increasingly used in recent years, allows to analyze relative polar and thermally unstable compounds without prior derivatization. Different types of MS instruments are available to fragment ionized molecules and detect their charged fragments providing structural information. Using these instruments, several MS/MS databases and software tools have been generated in order to facilitate metabolite identification (Horai, et al., 2010; Wishart, et al., 2009). However, most approaches are limited to identity searches, in which the unknown compound needs to be present in the database or to the assignment of unknowns into compound classes based on similar fragmentation (Rasche, et al., 2012; Sheldon, et al., 2009). This leaves the complete structure elucidation of unknowns still a major challenge.
Compared to MS/MS approaches in which fragments can be immediately fragmented further after generation, MS^n methods provide a more detailed fragmentation pathway. As only precursor ions are exited, fragments cool after formation, thus avoiding further fragmentation giving deeper insight into relationships between fragments (Stein, 2012). With the use of such a multidimensional mass spectrometry approach spectral trees can be generated containing precursor-product ion relationships (Sheldon, et al., 2009). Based on correlations between specific fragments in the MS^n spectra and the substructure of the measured molecule, structural elucidation of unknown compounds can be performed (Chen, et al., 2002; Fandino, et al., 2002; Wang, et al., 1999). However, as extensive databases and software tools are scarce, MS^n data handling and interpretation is mainly still limited to the manual identification of characteristic fragments from defined classes of compounds (Cui, et al., 1999; Kang, et al., 2007; Rochfort, et al., 2008), leaving the full power of multiple stage fragmentation mass spectra untapped. As a consequence our group recently extended the MEF (mass elemental formula) tool (Rojas-Cherto, et al., 2011; Rojas-Cherto, et al., 2012) for processing and comparing MS^n data which extracts ion-signals from MS^n spectra, followed by assignment of their corresponding elemental composition.
In previous work, the functionality of the MEF tool has been demonstrated for the determination of similarities between database entries (Rojas-Cherto, et al., 2012) and the tentative identification of small molecules (Peironcely, et al., 2013). Howe-

ver, the conserved structural information of the analyte, reflected by specific fragmentation mechanisms and the chemical structure of fragments and neutral-losses, was not exploited.

Here, we present a chemoinformatics guided pipeline for the structural identification of unknown compounds coined CMCP (Chemoinformatics supported MS^n Comparison Pipeline). By combining small scale sample extraction, fragmentation tree acquisition and the extended MEF tool with MS expert knowledge, the chemical structure of low concentrated complex molecules can be identified using solely mass spectrometry (Figure 1). We discuss the most important conceptual outcomes of a database query using CMCP and demonstrate how this information can ultimately be used to deduce complete chemical structures for unknown compounds.

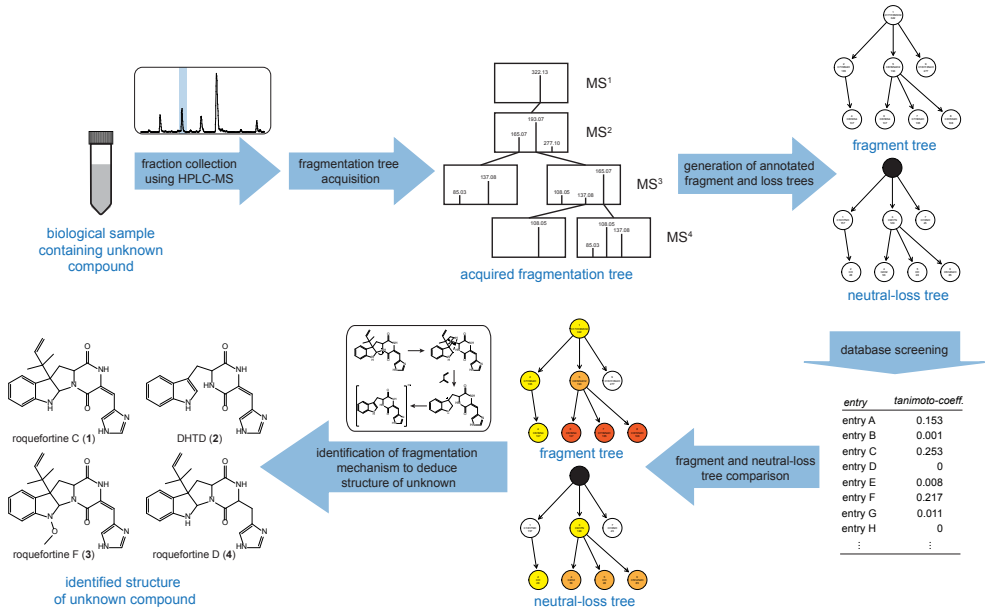


Figure 1. Schematic view of the CMCP pipeline for the identification of unknown compounds in a biological sample and structures of identified compounds **1** – **4**.

Experimental procedures

Strain and culture conditions

Penicillium chrysogenum strain DS54555, which lacks the *Ku70* and penicillin cluster genes was supplied by DSM Anti-infective (Delft, The Netherlands). Cells were grown on SMP medium (glucose, 5.0 g/L; lactose, 75 g/L; urea, 4.0 g/L; Na_2SO_4, 4.0 g/L; CH_3COONH_4, 5.0 g/L; K_2HPO_4, 2.12 g/L; KH_2PO_4, 5.1 g/L) using a shaking incubator at 200 rpm for 168 hours at 25°C.

Small scale isolation of analytes

Low amounts of the four structurally complex metabolites **1** – **4** (Figure 1), were ex-

tracted from liquid cultures of *P. chrysogenum* using the LC-UV-MS method described elsewhere (Ali, et al., 2013). Briefly, 230 µL of methanol was added to 50 µL of *P. chrysogenum* culture filtrate for protein precipitation. The sample was vortexed, centrifuged and 100 µL supernatant transferred to an Eppendorf vial which was subsequently evaporated in a Thermo-Speedvac (Thermo Scientific, San Jose, CA). The dried sample was re-dissolved in 100 µL water containing 2 % acetonitrile, vortexed and transferred to an autosampler vial. For separation, an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled to a Surveyor PDA detector (Thermo Scientific, San Jose, CA) and LTQ-FT Ultra mass spectrometer (Thermo Scientific, San Jose, CA) was used. A sample of 5 µL was injected onto a Waters Atlantis T3 column (2.1 x 100 mm, 3 µm) (Waters, Milford, MA). The elution was performed with a linear gradient from 98 % of solvent A (1 % acetonitrile and 0.1 % formic acid in water) and 2 % solvent B (1 % water and 0.1 % formic acid in acetonitrile) to 100 % of Solvent B in total of 25 minutes (first 1.5 minutes isocratic at 98% A, then to 40% of solvent B at 22 minutes and 10% solvent B at 25 minutes) at a flow rate of 300 µL/min. The column was flushed for 10 minutes at 100% B followed by equilibration for 8 minutes at 100 % A. Fractions of **1**, **2**, **3** and **4** were collected manually and subsequently evaporated to dryness in a Speedvac.

Preparation of Working Solutions
Working solutions were prepared by re-dissolving the dried fraction of each metabolite in 20 µL water containing 50% acetonitrile and 0.1% formic acid. Right before acquisition, working solutions were mixed with 5 µL isopropanol in a 384 well plate to increase nano-electrospray stability (Eppendorf, Hamburg, Germany).

Fragmentation tree acquisition, processing and database comparison
Fragmentation tree acquisition was carried out on a LTQ-Orbitrap-XL mass spectrometer (Thermo Fisher Scientific, Waltham, MA) equipped with an Advion Nanomate nano-electrospray source (Advion, Ithaca, NY) (1,5 kV source voltage, 120°C capillary temperature, 5 V capillary voltage, 70 V tube lens voltage and 35 % normalized collision energy for CID fragmentation) according a modified protocol described elsewhere (Kasper, et al., 2012). The scan rate of the mass spectrometer was set to 6 micro scans. At least three complete fragmentation trees were acquired per compound.
The ten most abundant fragment ions in MS^2, the nine highest in MS^3 and three highest in MS^4 were isolated and further fragmented resulting in 371 (1 MS^2, 10 MS^3, 90 MS^4 and 270 MS^5) possible fragmentations. Isolation width was set to 1.5 *m/z* and minimal precursor intensity required for further fragmentation was fixed to 50.000 counts. Acquired Thermo Xcalibur files were converted into mzXML format using ReadW software (Pedrioli, et al., 2004). Chemical formulas were assigned to all fragments of the acquired tree using the MEF tool (Rojas-Cherto, et al., 2011). MEF processing parameters were as following: Signal-to-noise ratio 1, mass accuracy 10 without applying the nitrogen rule or degree of unsaturation (Pellegrin, 1983). Finally, all processed trees were compared to an in-house metabolite MS^n database, containing more than different 400 entries, to find similar fragments and neutral-loss patterns (Rojas-Cherto, et al., 2012).

Results

Fragmentation tree generation and similarity search
To demonstrate the different conceptual outcomes of a database query and to show how this information can be used to deduce structural information, the *de novo* identification of the four unknown metabolites roquefortine C (**1**), dehydrohistidyltryptophanyldiketopiperazine (DHTD) (**2**), roquefortine F (**3**) and roquefortine D (**4**) is described in detail (Figure 1). Compounds **1** - **4** were obtained, next to six additional compounds, from comparative metabolite profiling of host and various deletion strains of *P. chrysogenum*, using high resolution mass spectrometry (HPLC-HR-MS) (Ali, et al., 2013). Their mass was determined as 390.192 (**1**), 322.129 (**2**), 420.203 (**3**) and 392.208 (**4**) dalton, corresponding to the protonated molecules with formula $C_{22}H_{24}N_5O_2$, $C_{17}H_{16}N_5O_2$, $C_{23}H_{26}N_5O_3$ and $C_{22}H_{26}N_5O_2$. As none of the ions had a corresponding entry in a public MS database like MassBank (Horai, et al., 2010), HMDB (Wishart, et al., 2009) or PRIMe (Akiyama, et al., 2008; Sakurai, et al., 2013) and because the biosynthetic mechanism of their producing gene cluster was not known, additional structural information was not available leaving their structure completely unknown. In order to acquire multiple stage fragmentation mass spectra, small scale fraction collection was performed to extract low amounts of pure analytes from *P. chrysogenum* culture broth. Subsequently, fragmentation trees were acquired and processed according the description in the method section. The resulting fragment and neutral-loss trees were compared to an in-house metabolite MS^n database, which was searched for similar mass fragments and neutral-losses present in database entries. Database entries were ranked according their degree of similarity, represented by the Tanimoto coefficient (Flower, 1998), yielding the database entry of protonated roquefortine C with formula $C_{22}H_{24}N_5O_2$ as most similar compound to all four metabolites.

Structure elucidation based on identical fragment and neutral-loss trees
With the precursor and almost all fragments and neutral-losses of **1** being present and similarly arranged as in the fragment and neutral-loss tree of roquefortine C, an identical chemical structure of **1** and roquefortine C was indicated (Figure 2). This was supported by a comparison of their multiple stage fragmentation spectra which showed fragments with almost identical mass-over-charge ratios and relative intensities over multiple fragmentation stages (data not shown). Based on the known structure of the database entry, the structure of **1** was tentatively assigned as roquefortine C which was ultimately confirmed using HPLC-MS/MS comparing the retention time of **1** and a roquefortine C standard

Structure elucidation based on similar fragment and neutral-loss trees
For compound **2**, the database query returned partial similarity to its most similar database entry roquefortine C with almost all fragments and neutral-losses of compound **2** present and similarly arranged as in the database entry (Figure 3, Supplemental Figure 1 and 2). However, compared to the identification of compound **1**, in which compound **1** and the database entry showed the same ion in MS^1, here different ions were observed, respectively *m/z* 322 (ID 1, MS^1, $C_{17}H_{16}N_5O_2$) in compound **2** and *m/z* 390 (ID 1, MS^1, $C_{22}H_{24}N_5O_2$) in roquefortine C.

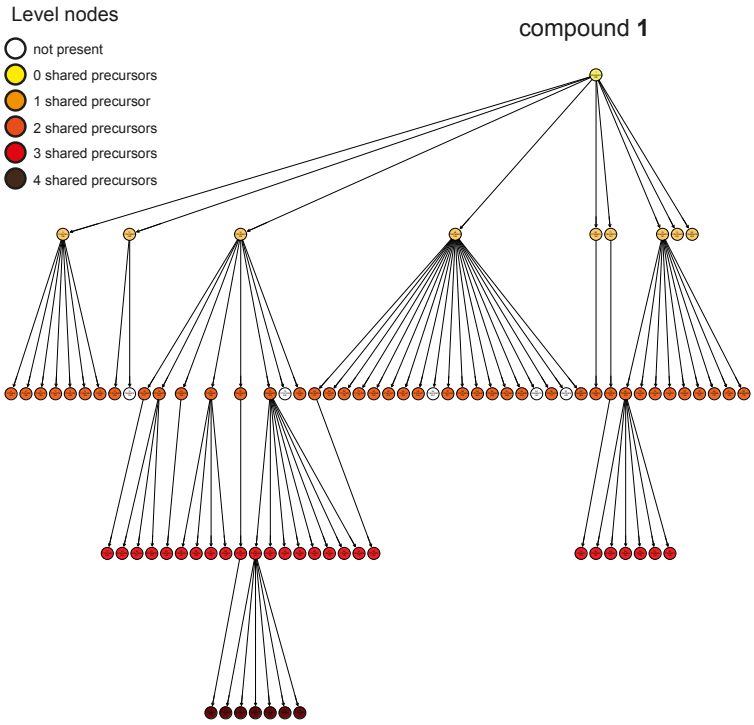


Figure 2. Fragment tree of compound **1** after comparison to the fragment tree of the most similar database entry roquefortine C. Fragments present in both compounds are colored according their position in a shared fragmentation path. For the visualization of similarities overlapping nodes present in both fragment trees are colored, unique ions present only in one tree are white. Different colors indicate the length of the largest path of consecutive fragments that two fragment trees have in common. A comprehensive description of the fragmentation tree representation was published earlier (Rojas-Cherto, et al., 2012).

For further structure elucidation, the two highly similar subtrees of fragment *m/z* 322 (ID 28, MS^2, $C_{17}H_{16}N_5O_2$) in roquefortine C and *m/z* 322 (ID 1, MS^1, $C_{17}H_{16}N_5O_2$) in compound **2** were used. Their similarly arranged fragments and neutral-losses, containing an identical elemental composition, indicated that the fragment with *m/z* 322 originating from the precursor ion *m/z* 390 (ID 1, MS^1, $C_{22}H_{24}N_5O_2$) in roquefortine C has an identical structure as the ion *m/z* 322 of **2**, which represents the protonated molecule in MS^1. This was further supported by comparing the multiple stage fragmentation spectra of their common subtrees which showed fragments with almost identical mass-over-charge ratios and relative intensity over multiple fragmentation stages (Figure 4). By identifying the chemical structure of the fragment *m/z* 322 in roquefortine C, formed by a loss of isoprene, and transferring the information to **2**, the structure of protonated **2** could be determined as dehydrohistidyltryptophanyldiketopiperazine (DHTD) (Figure 5). Its tentative structure was ultimately confirmed by NMR experiments (Ali, et al., 2013).

An example in which the fragmentation tree of a database entry is a complete subtree of an unknown compound is the identification of compound **3**, which shares a common subtree with its most similar database entry roquefortine C, initiated by

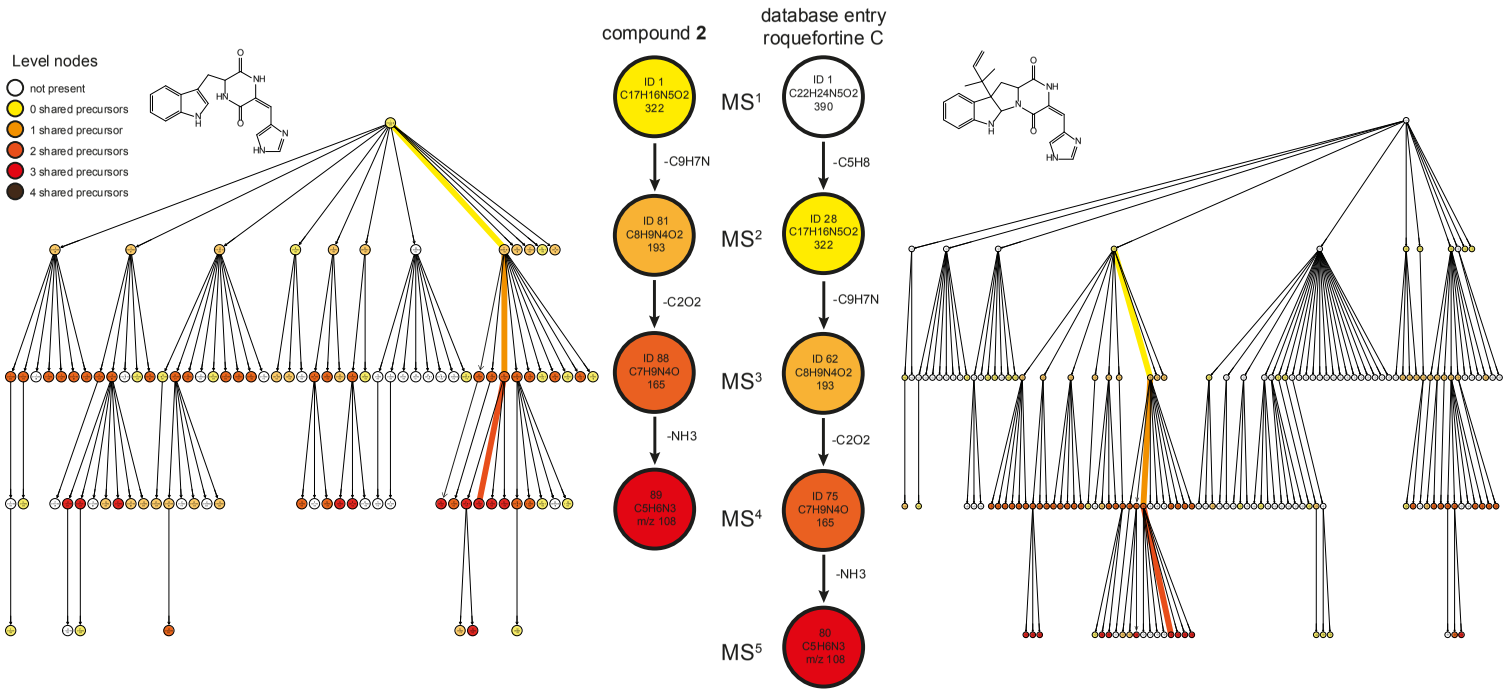


Figure 3. Fragment tree comparison of compound **2** (left) and its most similar database entry roquefortine C (right) revealed that the entire tree of **2** is a subtree of roquefortine C. The shared fragmentation path of the most abundant fragment per spectrum is highlighted by colored arrows and shown in more detail in the middle of the figure.

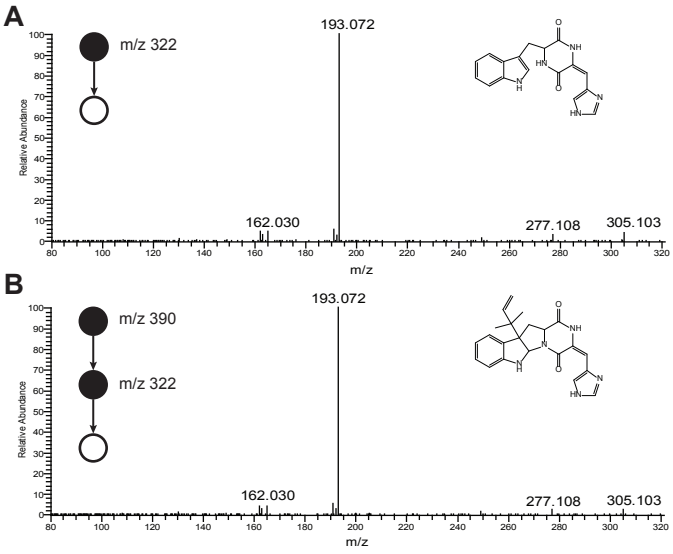


Figure 4. MS^2 fragmentation spectrum of compound **2** (A) and MS^3 fragmentation spectrum of the database entry roquefortine C (B) showing an identical fragmentation pattern.

ion *m/z* 390 (ID 1, MS^1, $C_{22}H_{24}N_5O_2$) in roquefortine C and *m/z* 390 (ID 100, MS^2, $C_{22}H_{24}N_5O_2$) in **3** (Supplemental Figure 3 and 4). As the complete tree of roquefortine C is present in the tree of **3**, including the precursor ion in MS^1, an identical structure for the ion *m/z* 390 in both compounds is concluded. This is supported by identical fragmentation spectra over multiple fragmentation stages initiated by a fragmentation of the ions at *m/z* 390 in both compounds (Supplemental figure 5). Compared to roquefortine C, compound **3** contains one additional carbon, two additional hydrogens and one additional oxygen which were cleaved off during first fragmentation yielding ion *m/z* 390 (ID 100, MS^2, $C_{22}H_{24}N_5O_2$) with an identical structure as protonated roquefortine C. Supported by the competing losses of $CH_3O\bullet$ and CH_4O from a fragmentation of *m/z* 420 (ID 1, MS^1, $C_{23}H_{26}N_5O_3$) yielding the fragments *m/z* 389 (ID 46, MS^2, $C_{22}H_{23}N_5O_2$) and *m/z* 388 (ID 29, MS^2, $C_{22}H_{22}N_5O_2$), the structure of the loss producing *m/z* 390 could be determined as formaldehyde originating from the cleavage of a methoxygroup (Stevigny, et al., 2004). Therefore, a roquefortine C structure is concluded for compound **3** with a methoxygroup attached. Its position in the structure of **3** is unclear, as the methoxygroup is cleaved off in the first

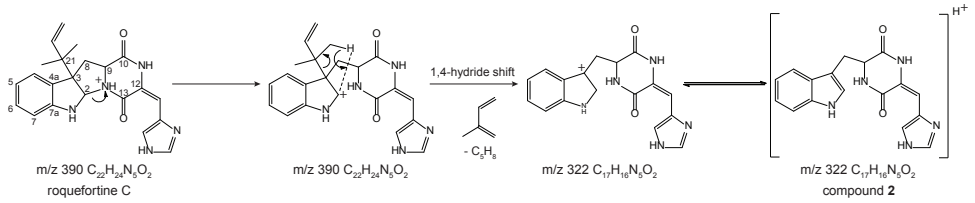


Figure 5. Proposed fragmentation mechanism of database entry roquefortine C yielding compound **2** *in situ* in the mass spectrometer. After an initial protonation, the bond between C2 and N14 is cleaved in roquefortine C followed by a 1,4-hydride shift leading to the loss of isoprene, yielding *m/z* 322.

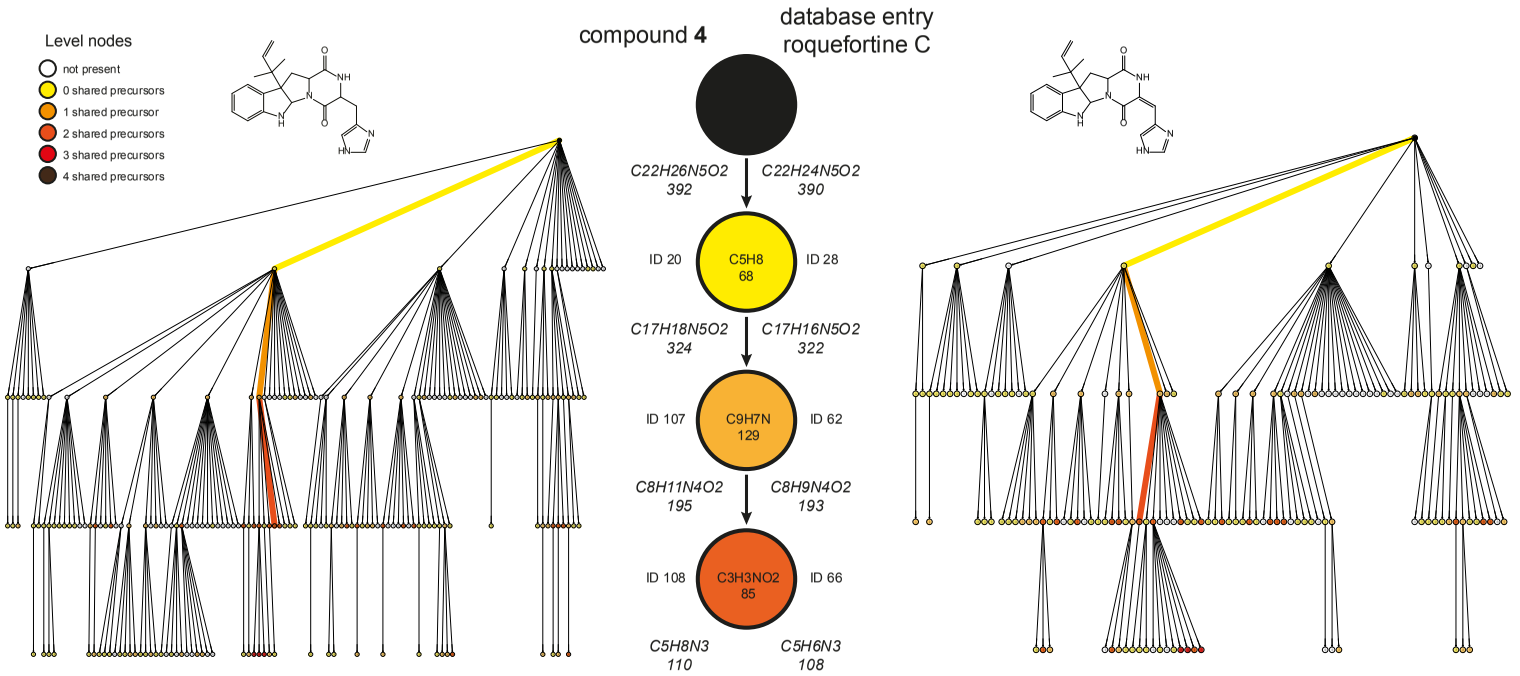


Figure 6. Neutral-loss tree comparison of compound **4** (left) and roquefortine C (right) showing highly similar neutral-loss cascades present in both trees. The shared neutral-loss path used for the identification of **4** is shown in the middel in more detail. Detailed neutral-loss trees of roquefortine C and **4** obtained from neutral-loss tree comparison can be found in the supplement.

fragmentation stage. Using NMR experiments, the structure of **3** was determined as roquefortine F, containing the structure of roquefortine C with a methoxygroup on the nitrogen in the indol part of the molecule which is in good agreement with the results obtained from CMCP (**Chapter 3**).
It should be pointed out that although the protonated form of **2** and roquefortine C share the identical ion *m/z* 322, their subtrees differ significantly in depth and width. Therefore, not all fragments of *m/z* 322 (ID1, MS^1) in **2** could be found in the subtree of *m/z* 322 (ID28, MS^2) in roquefortine C. This is due to the fact that next to different observed absolute intensities of the ions, two ions originating from different fragmentation stages are compared. While for **2**, ion *m/z* 322 was found in MS^1, for roquefortine C it was found in MS^2. Due to the limitations of the acquisition protocol, the further fragmentation of an ion in MS^1 can result in maximal 371 fragmentations (1 MS^2, 10 MS^3, 90 MS^4, 270 MS^5) whereas the further fragmentation for an ion in MS^2 is limited to maximal 37 fragmentations (1 MS^3, 9 MS^4, 27 MS^5). As a consequence, a much richer fragmentation tree was acquired for the ion *m/z* 322 in compound **2**. Similar results were obtained for the identical ion *m/z* 390 present in MS^1 of roquefortine C and MS^2 in **3** for which a much a richer fragmentation was observed in roquefortine C.

Structure elucidation based on similar neutral-loss cascades
For compound **4**, the database query returned partial similarity to its most similar database entry roquefortine C, showing similar neutral-loss cascades over various fragmentation stages with only few shared unconnected fragments (Figure 6, Figure 7, Supplemental Figure 6 and 7). For further structure identification, the neutral-loss cascade C_5H_8 – C_9H_7N – $C_3H_3NO_2$ shared by roquefortine C (ID 28, ID 62, ID 66) and compound **4** (ID 20, ID 107, ID 108) was used. As this relative specific neutral-loss cascade is shared by both compounds, it can be concluded that **4** might have similar

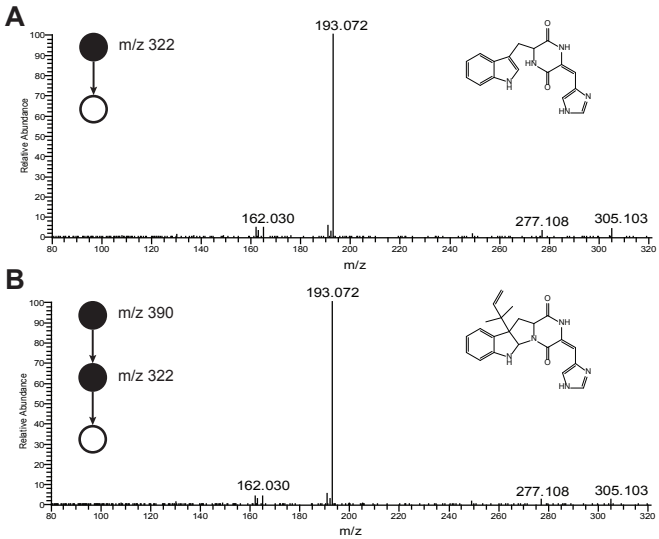


Figure 4. MS^2 fragmentation spectrum of compound **2** (A) and MS^3 fragmentation spectrum of the database entry roquefortine C (B) showing an identical fragmentation pattern.

structural features as roquefortine C, resulting in similar fragmentation mechanisms with similar neutral-losses upon excitation. Identifying these common fragmentation mechanisms, based on the known structure of the database entry, resulted in a partial identification of **4**, containing a dimethylallyl (C_5H_8), indol (C_9H_7N) and diketopiperazine ($C_3H_3NO_2$) moiety similarly arranged (Figure 8). Striking is to see that the difference in the elemental composition between both compounds, two additional hydrogens in the structure of **4**, are conserved throughout the entire fragmentation process and still being present in the final fragment with *m/z* 110 (ID 108, MS^4, $C_5H_8N_3$). Although the structure of corresponding fragment of roquefortine C can be deduced based on the known structure of roquefortine C, structural information about the final fragment of **4** is not available, as a further fragmentation was not successful. However, in combination with gene deletion experiments (Ali, et al., 2013), concluding that compounds **1 – 4** originate from the same biosynthetic gene cluster, the structure of **4** was tentatively identified as roquefortine D which was subsequently confirmed by NMR experiments (Ali, et al., 2013).

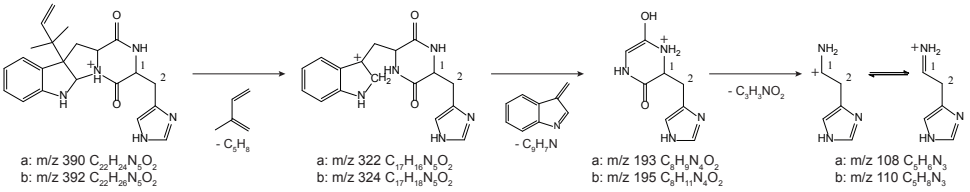


Figure 8. Proposed fragmentation of database entry roquefortine C (a) and compound **4** (b). After protonation, isoprene, represented by the loss of C_5H_8, is cleaved off in the first fragmentation step yielding the fragment *m/z* 322 in roquefortine C and *m/z* 324 for compound **4**. Due to the subsequent loss of the indole part of the structure, indicated by a loss of C_9H_7N, ions *m/z* 193 in roquefortine C and *m/z* 195 in **4** are formed. In the final fragmentation step the ring of the diketopiperazine structure opens, leading to a loss of $C_3H_3NO_2$ and the formation of *m/z* 108 in roquefortine C and *m/z* 110 in **4**.

Discussion

The results presented here demonstrate how the metabolite identification pipeline CMCP can be used to obtain detailed structural information of structurally complex molecules. Using fragment and neutral-loss tree comparison, structural similar compounds were extracted from a MS^n database. As structural information is coded in the fragments, neutral-losses and fragmentation mechanisms, structural similarities between the unknown compound and the database entry are revealed. Identifying these shared mechanisms, linked to substructure information stored in the database or using MS expert knowledge, and applying them to the unknown molecule allows complete (sub)structures identification even without the unknown being present in the database, as shown for the *de novo* identification of compounds **2** - **4**.

Depending on the structural overlap between the analyte and a database entry, high similarity can be a result of similar fragments, similar neutral-losses or a combination of both. Particularly the comparison of neutral loss trees, which is used here for the first time to obtain detailed structural information, represents a valu-

able structure elucidation approach as it uses complementary information to fragment trees for the determination of similarity. As almost no fragments are shared between the database entry roquefortine C and compound **4**, a comparison solely based on similar fragments would not have been as successful. However, it should be pointed out that there can be a large gap between the degree of similarity obtained from fragment tree comparison and neutral-loss tree comparison. This is due to the fact that similarity based on fragments can be verified as fragments can be further fragmented generating a fragmentation subtree which contains structural information about the particular fragment. In case the obtained multiple stage fragmentation spectra are identical, it can be assumed that their precursor possess an identical chemical structure. In contrast, loss tree comparison is based exclusively on uncharged neutral-losses which can, by definition, not further be fragmented. Although their structures can be deduced from the known structure of the database entry and applied to the unknown analyte, as shown for the identification of compound **4**, it is not possible to absolutely confirm the identical identity of shared losses in both compounds. Certainty that the exact same losses with the same fragmentation mechanisms occur in both compounds is depending on the length of the shared loss path and the uniqueness of the losses. The loss path C_5H_8 – C_9H_7N – $C_3H_3NO_2$ used for the identification of **4** contains relatively unique fragments corresponding to dimethylallyl, tryptophan and diketopiperazine substructures which are only shared with the database entry roquefortine C.

Uncovering relationships between neutral-losses and providing extensive structural information about individual fragments by further fragmenting them, is the main advantage of MS^n methods compared to MS/MS techniques, as it offers crucial information beyond elemental compositions and intensities. However, the structural information conserved in the fragments also needs to be exploited. Especially for smaller databases which do not provide enough positive hits to determine the structure of a fragment or neutral-loss solely on the structural overlap of a common substructure of multiple similar database entries, fragments and losses need to be manually assigned to a particular part of the molecule. Without using this structural information, the identification of **2** - **4** would have been limited to a structural classification resulting in a somehow relationship to roquefortine C. With the advent of novel software tools for the processing of multiple stage fragmentation data, tandem mass spectra databases may be replaced by MS^n databases due to the additional structural information available within.

Traditionally, the number of analytes which can be identified using mass spectrometry depends on the comprehensiveness of the database, as identity searches require each analyte to be represented by an identical database entry. Thus, large databases with numerous entries are necessary to cover a large chemical space. By searching for fragmentation similarities between multiple stage fragmentation trees as presented here, rather than complete identities, less comprehensive databases are required. For the identification of compounds **1** – **4**, only one database entry was necessary to represent the fragmentation of a specific structural group of compounds, as shown for roquefortine C which was used to identify the metabolites **1** – **4**.

In addition, CMCP paves the way towards a fully automated structure identification using solely mass spectrometry. With the ongoing storage of structural information

of fragments and neutral-losses into the MS^n database, detailed structural information of nodes in database entries is extended. Using this information, structures of similar nodes present in the unknown target compound can be automatically assignment enabling an immediate identification of fragments and neutral-losses without much intervention of MS experts. By automatically combining the structural overlap of various subtrees or losses even from different similar database entries, complete (sub)structures could be deduced allowing a completely automated structure identification.

Conclusion

Here, we present a structure elucidation pipeline that allows the complete structural identification of unknown molecules beyond structural classification without the need of the compound to be present in the database. It could be shown that fragment and neutral-loss tree comparison yield complementary information which makes neutral-loss tree comparison a valuable addition for the identification of unknowns. In combination with the ongoing structural identification of fragments and neutral-losses, automatic substructure assignment is possible, paving the way for a completely automated structure elucidation.
Concluding, fragmentation tree comparison is an important component of our metabolite identification pipeline which combines comparative metabolite profiling and structure elucidation using CMCP and NMR analysis. With this pipeline various (novel) metabolites of *P. chrysogenum* were identified and their biosynthesis could be determined. Future work on direct-infusion experiments and data processing tools aim to improve the throughput of this pipeline by overcoming the relative resource-intensive small scale extraction of the unknown compounds.

Acknowledgements

This project was supported by the Perspective Genbiotics program subsidized by the Stichting voor de Technische Wetenschappen (STW) and (co)financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. The authors are grateful to Prof. Dr. Nico M. Nibbering for his help and guidance during the elucidation of fragmentation mechanisms.

References

Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., Hirai, M.Y., Sakurai, T., Kikuchi, J., and Saito, K. (2008). PRIMe: a Web site that assembles tools for metabolomics and transcriptomics. In Silico Biol 8, 339-345.

Ali, H., Ries, M.I., Nijland, J.G., Lankhorst, P.P., Hankemeier, T., Bovenberg, R.A., Vreeken, R.J., and Driessen, A.J. (2013). A Branched Biosynthetic Pathway Is Involved in Production of Roquefortine and Related Compounds in Penicillium chrysogenum. PloS one 8, e65328.

Chen, G., Pramanik, B.N., Bartner, P.L., Saksena, A.K., and Gross, M.L. (2002). Multiple-stage mass spectrometric analysis of complex oligosaccharide antibiotics (everninomicins) in a quadrupole ion trap. J Am Soc Mass Spectrom 13, 1313-1321.

Cui, M., Sun, W., Song, F., Liu, Z., and Liu, S. (1999). Multi-stage mass spectrometric studies of triterpenoid saponins

in crude extracts from Acanthopanax senticosus harms. Rapid communications in mass spectrometry : RCM 13, 873-879.

Fandino, A.S., Karas, M., Toennes, S.W., and Kauert, G. (2002). Identification of anhydroecgonine methyl ester N-oxide, a new metabolite of anhydroecgonine methyl ester, using electrospray mass spectrometry. J Mass Spectrom 37, 525-532.

Flower, D.R. (1998). On the Properties of Bit String-Based Measures of Chemical Similarity. J. Chem. Inf. Comput. Sci 38, 379-386.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45, 703-714.

Kang, J., Hick, L.A., and Price, W.E. (2007). A fragmentation study of isoflavones in negative electrospray ionization by MSn ion trap mass spectrometry and triple quadrupole mass spectrometry. Rapid Commun Mass Spectrom 21, 857-868.

Kasper, P.T., Rojas-Cherto, M., Mistrik, R., Reijmers, T., Hankemeier, T., and Vreeken, R.J. (2012). Fragmentation trees for the structural characterisation of metabolites. Rapid Commun Mass Spectrom 26, 2275-2286.

Kind, T., and Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC bioinformatics 7, 234.

Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., et al. (2004). A common open representation of mass spectrometry data and its application to proteomics research. Nature biotechnology 22, 1459-1466.

Peironcely, J.E., Rojas-Cherto, M., Tas, A., Vreeken, R., Reijmers, T., Coulier, L., and Hankemeier, T. (2013). Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics. Analytical chemistry 85, 3576-3583.

Pellegrin, V. (1983). Molecular formulas of organic compounds: the nitrogen rule and degree of unsaturation. J. Chem. Educ. 60, 626.

Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatos, A., and Bocker, S. (2012). Identifying the unknowns by aligning fragmentation trees. Analytical chemistry 84, 3417-3426.

Rochfort, S.J., Trenerry, V.C., Imsic, M., Panozzo, J., and Jones, R. (2008). Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. Phytochemistry 69, 1671-1679.

Rojas-Cherto, M., Kasper, P.T., Willighagen, E.L., Vreeken, R.J., Hankemeier, T., and Reijmers, T.H. (2011). Elemental composition determination based on MS(n). Bioinformatics 27, 2376-2383.

Rojas-Cherto, M., Peironcely, J.E., Kasper, P.T., van der Hooft, J.J., de Vos, R.C., Vreeken, R., Hankemeier, T., and Reijmers, T. (2012). Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. Analytical chemistry 84, 5524-5534.

Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K., Hirai, M.Y., and Saito, K. (2013). PRIMe Update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. Plant Cell Physiol 54, e5.

Sheldon, M.T., Mistrik, R., and Croley, T.R. (2009). Determination of ion structures in structurally related compounds using precursor ion fingerprinting. J Am Soc Mass Spectrom 20, 370-376.

Stein, S. (2012). Mass spectral reference libraries: an ever-expanding resource for chemical identification. Analytical chemistry 84, 7274-7282.

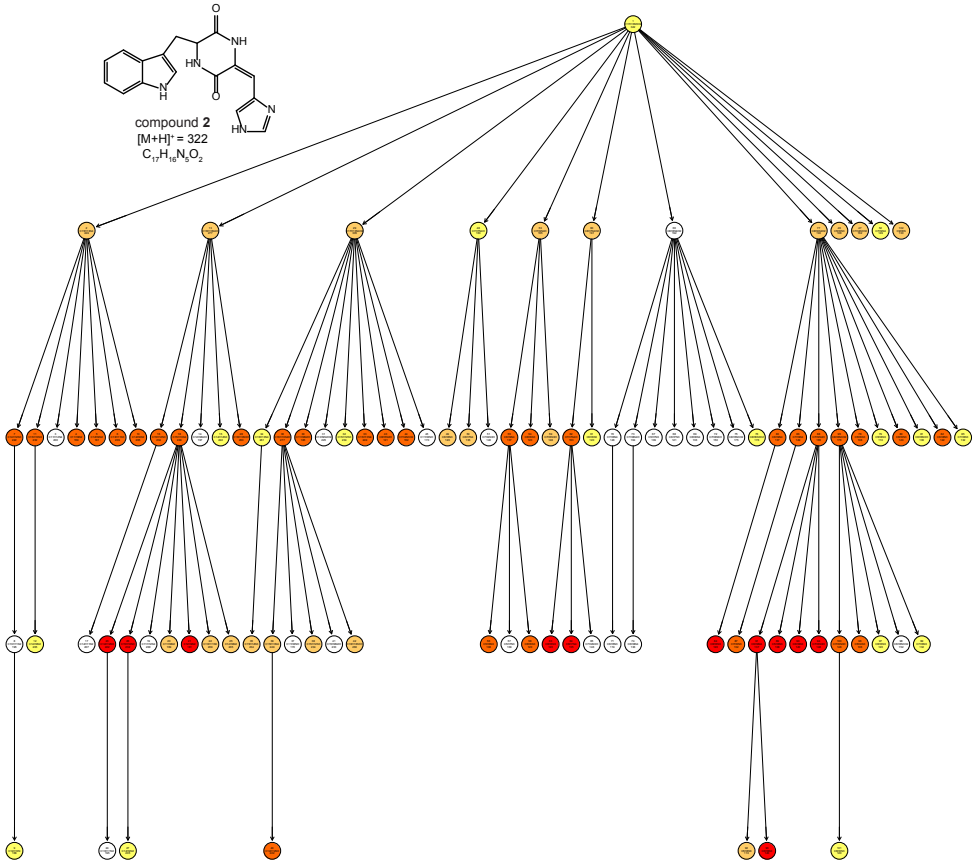
Stevigny, C., Jiwan, J.L., Rozenberg, R., de Hoffmann, E., and Quetin-Leclercq, J. (2004). Key fragmentation patterns of aporphine alkaloids by electrospray ionization with multistage mass spectrometry. Rapid Commun Mass Spectrom 18, 523-528.

Villas-Boas, S.G., Mas, S., Akesson, M., Smedsgaard, J., and Nielsen, J. (2005). Mass spectrometry in metabolome analysis. Mass Spectrom Rev 24, 613-646.

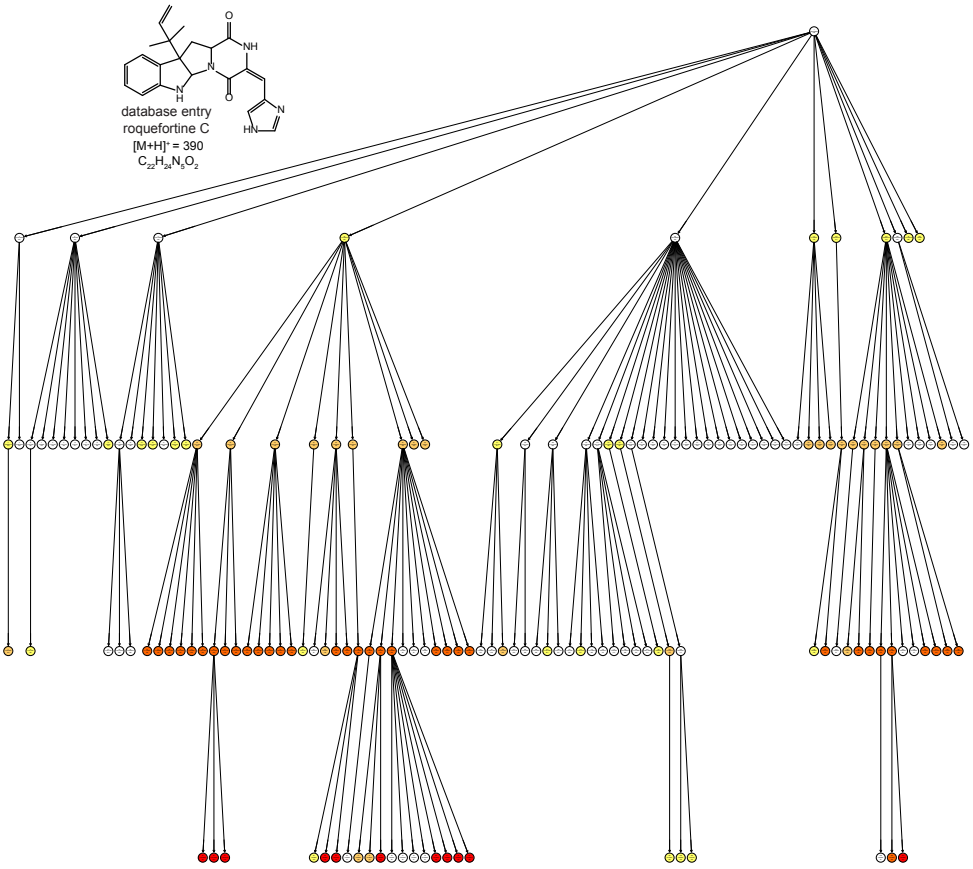
Wang, W., Liu, Z., Ma, L., Hao, C., Liu, S., Voinov, V.G., and Kalinovskaya, N.I. (1999). Electrospray ionization multiple-stage tandem mass spectrometric analysis of diglycosyldiacylglycerol glycolipids from the bacteria Bacillus pumilus. Rapid Commun Mass Spectrom 13, 1189-1196.

Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., et al. (2009). HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37, D603-610.

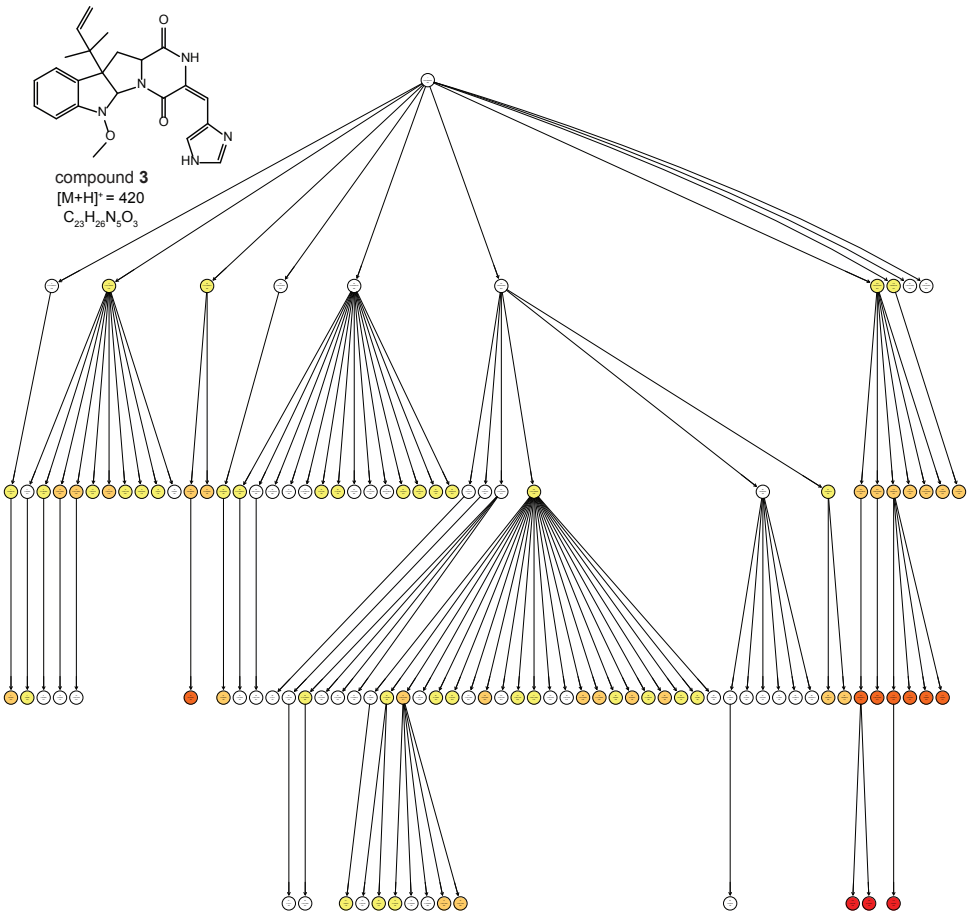
Supplemental Information



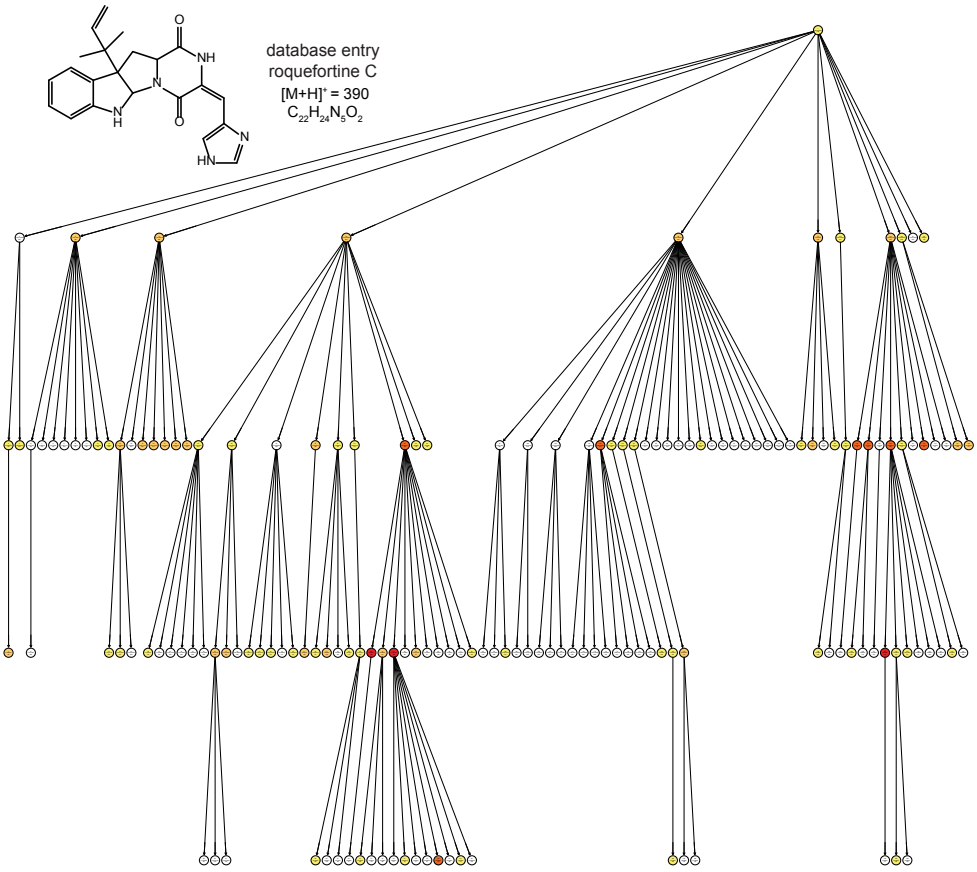
Supplemental Figure 1. Fragment tree of compound **2** compared to the fragment tree of database entry roquefortine C as indicated by colored nodes.



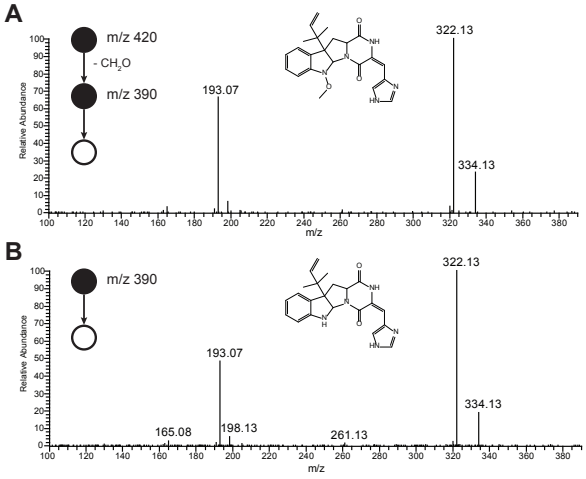
Supplemental Figure 2. Fragment tree of the database entry roquefortine C compared to the fragment tree of compound **2**.



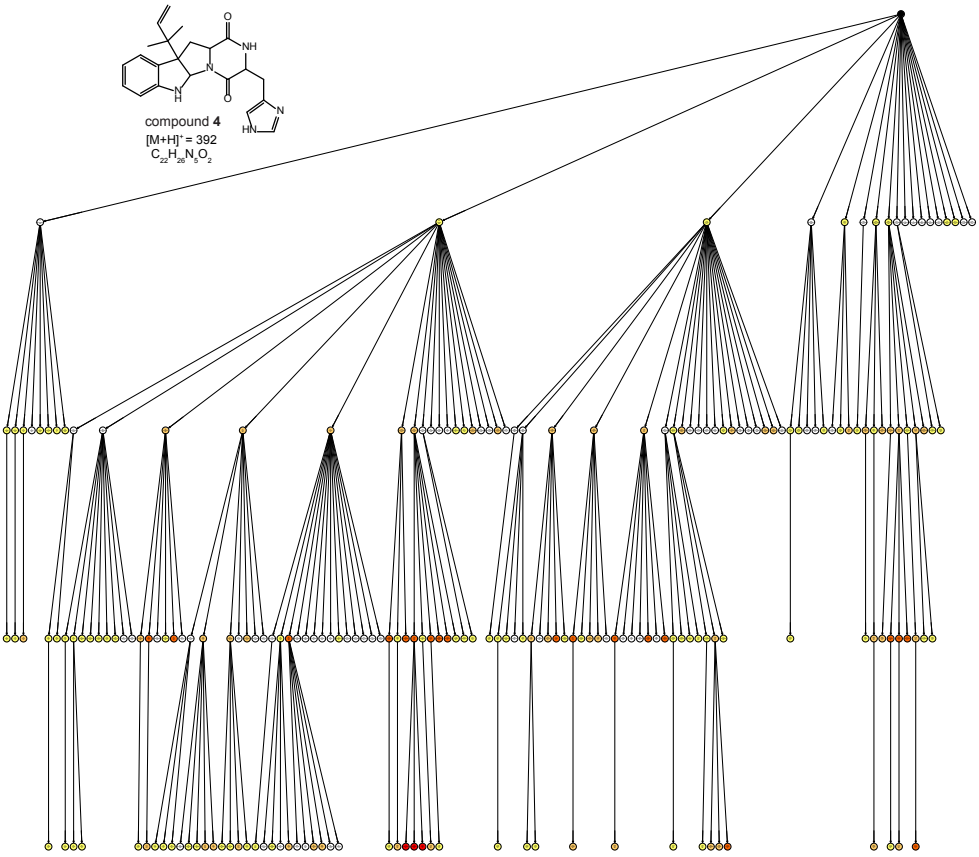
Supplemental Figure 3. Fragment tree of compound **3** compared to the fragment tree of database entry roquefortine C.



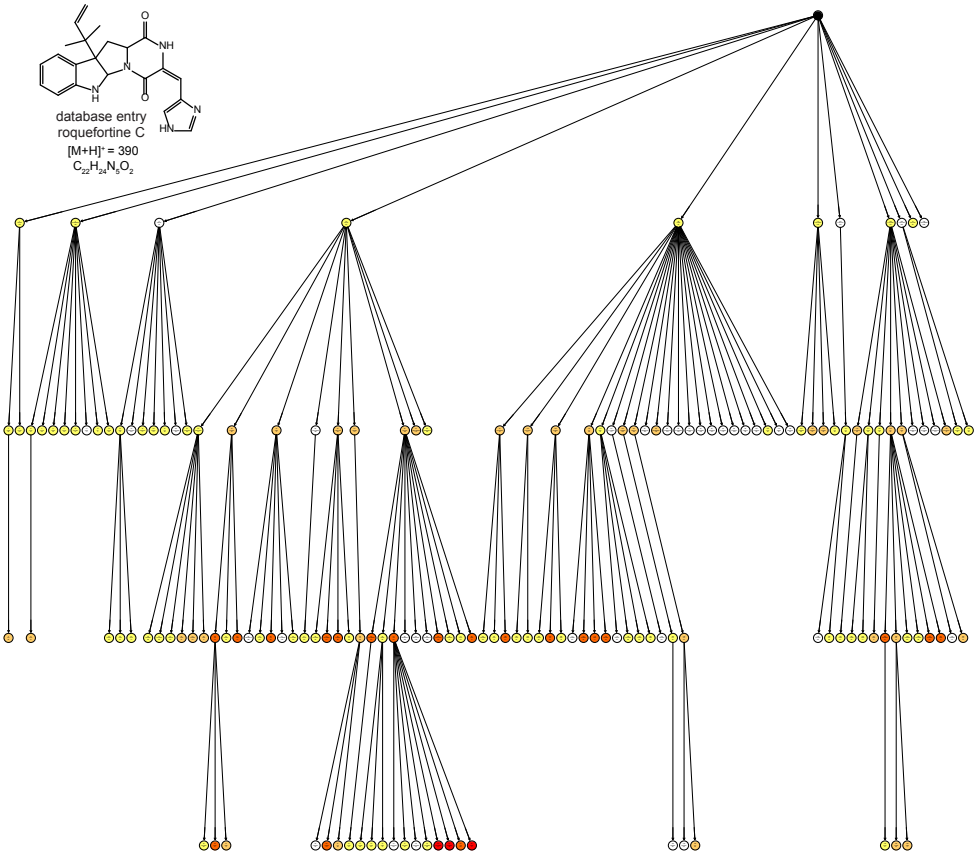
Supplemental Figure 4. Fragment tree of the database entry roquefortine C compared to the fragment tree of compound **3**.



Supplemental Figure 5. MS^3 fragmentation spectrum of compound **3** (A) and MS^2 fragmentation spectrum of the database entry roquefortine C (B) showing an identical fragmentation pattern.



Supplemental Figure 6. Neural-loss tree of compound **4** compared to the neutral-loss tree of the database entry roquefortine C.



Supplemental Figure 7. Neutral-loss tree of database entry roquefortine C compared to neutral-loss tree of compound **4**.

Chapter 6

Multiple stage fragmentation tree comparison enables detailed structure elucidation in direct infusion mass spectrometry based experiments

Based on

Marco I. Ries, Hazrat Ali, Arnold J.M. Driessen, Thomas Hankemeier, Theo Reijmers, Rob J. Vreeken
Multiple stage fragmentation tree comparison enables detailed structure elucidation in direct infusion mass spectrometry based experiments
In preparation for publication

Abstract

Direct infusion mass spectrometry (DI-MS) is a fast, reliable, sensitive and cost effective method which is particularly attractive when dealing with large sample sets. However, structure identification in direct infusion based experiments is still a major obstacle due to co-fragmenting interferences like isobaric and isomeric ions, limiting most applications to metabolite fingerprinting only.
Here we present the application of CMCP (Chemoinformatics supported MS^n Comparison Pipeline) for the identification of metabolites in direct infusion based experiments. We demonstrate the structure elucidation of secondary metabolites in a complex liquid culture of *Penicillium chrysogenum* using multiple stage fragmentation spectra from single precursor ions as well as with isobaric and isomeric interferences. Next to the structure elucidation with the sample compound being present in the database, structure identification based on similar structures is described enabling *de novo* structure elucidation in direct infusion based experiment beyond a classification of compounds into compound classes.
This study presents a potential method to identify unknown metabolites conveniently in DI, without additional instrumental configurations, extending DI to more than just a first line approach.

Introduction

With the prevalent application of metabolomics, several methods and techniques for the simultaneous analysis of a large number of compounds have been developed which provide detailed information about the metabolome of complex biological samples. Traditional approaches for the determination of metabolites by mass spectrometry use chromatographic separation like GC (Gas Chromatography), LC (Liquid Chromatography) or CE (Capillary Electrophoresis) before metabolite detection, thus generating qualitative or quantitative information on individual analytes (Dettmer, et al., 2007). These methods require laborious and time consuming sample preparation procedures and significant analysis time due to chromatographic separation. Furthermore, separation and detection of metabolites is heavily depending on the chemical nature of the analytes and the used chromatographic method. In general, a combination of various analytical techniques like HILIC, RP or Ion-pairing, targeting different classes of metabolites, is necessary as no single analytical platform allows a comprehensive analysis of the metabolome.
An alternative high-throughput approach to capture information related to the total metabolite content is the use of direct infusion mass spectrometry (DI-MS) (Boernsen, et al., 2005; Goodacre, et al., 2002; Koulman, et al., 2007) which renounces the limiting separation step before sample ionization, allowing the fast detection of various metabolite classes. During the ionization process of electrospray ionization (ESI), the most common method used for ionization in mass spectrometry based experiments, analytes become charged by the loss or gain of a proton, or other adducts. As minimal fragmentation takes place during ionization, the measured mass of an analyte, recorded with high mass accuracy and resolution, is expected to be close to the anticipated mass of a specific metabolite recorded in a database, potentially allowing a direct putative identification (Mungur, et al., 2005; Nakamura, et al., 2007; Raterink, et al., 2013). Several databases containing metabolite information are available in order to facilitate putative annotation of accurate mass signals (e.g. PubChem (Li, et al., 2010), HMDB (Wishart, et al., 2009), KEGG (Kanehisa, et al., 2012), KNapSAcK (Afendi, et al., 2012)). However, an accurate mass does not give any structural information beyond the molecular formula, limiting most direct infusion experiments to screening purposes only (Draper, et al., 2013). Using MS/MS, individual precursor ions can be selected and fragmented, resulting in specific fragment ions which can help to elucidate the structure of the molecule. The interpretation of MS/MS experiments is challenging as isobaric ions (same nominal mass) and isomeric ions (same chemical composition) are fragmented simultaneously, resulting in MS^2 spectra with mixed fragments of various precursor ions (Wichitnithad, et al., 2010). Although newer generation mass spectrometers substantially reduce co-fragmentations with isolation widths as narrow as 0.5 *m/z* (mass-to-charge ratio) (Savitski, et al., 2011), co-fragmentation of interferences can't be completely avoided leaving structure identification in direct infusion a major challenge. Therefore, more time consuming methods are currently required for a full structure elu-

cidation involving chromatographic separation and MS or NMR based approaches which in return, compromise the advantages of DI experiments. Especially finding back the target molecules, obtained from DI screening, in chromatographic based experiments is challenging as the physiochemical properties of the targeted analyte are mostly unknown.
In contrast to MS/MS approaches, multiple stage mass spectrometry (MS^n) enables a further fragmentation of fragments generating so called fragmentation trees. As solely selected ions are further fragmented, insight into multidimensional precursor-parent ion relationships is obtained. With the chemical structure of the precursor coded in its fragments and neutral-losses, structural information of the precursor ion can be deduced (Cui, et al., 1999; Kang, et al., 2007; Rochfort, et al., 2008). Recently, our group developed CMCP for the structure elucidation of compounds using solely multiple stage fragmentation tree comparison (**Chapter 5**). By comparing fragment and neutral-loss trees, compounds with similar fragmentation mechanisms can be extracted from a MS^n database (Kasper, et al., 2012; Rojas-Cherto, et al., 2012). Based on the identified shared fragmentation mechanisms, chemical structures of unknowns can be deduced without them being present in a database.

Here we show the enormous potential of CMCP for the structure elucidation of compounds from complex analyte mixtures using direct infusion. To demonstrate the capabilities of our approach to identify compounds from fragmentation data of one population of structurally identical precursors as well as from co-fragmented isobaric and isomeric ions, the structure elucidation of various complex metabolites from liquid cultures of the fungus *Penicillium chrysogenum* is described.

Experimental procedures

Material and Chemicals
3-me-7-(3-methylbenzyl)-8-((1-phenylethyl)amino)-3,7-dihydro-1H-purine-2,6-dione (**2**) was purchased from Sigma-Aldrich.

*Preparation of working and stock solutions of **1** and **2***
Stock solutions consist of 1 µg/mL analyte in ethanol. A working solution was prepared by mixing the stock solutions of compound **1** and **2** in a 1:5 ratio. Right before acquisition 10 µL working solution was mixed with 5 µL isopropanol and 10 µL water containing 0.1 % formic acid in a 384 well plate (Eppendorf, Hamburg, Germany).

Strain and culture conditions
P. chrysogenum strain DS54555, which lacks the *Ku70* and penicillin cluster genes was kindly supplied by DSM Anti-infective (Delft, The Netherlands). Cells were grown on SMP medium (glucose, 5.0 g/L; lactose, 75 g/L; urea, 4.0 g/L; Na_2SO_4, 4.0 g/L; CH_3COONH_4, 5.0 g/L; K_2HPO_4, 2.12 g/L; KH_2PO_4, 5.1 g/L) using a shaking incubator at 200 rpm for 168 hours at 25°C.

Sample preparation
To 50 µL of a thawed fermentation broth 8 µL internal standard mixture containing 855 nmol/mL ranitidine, 657 nmol/mL reserpine and 1144 nmol/mL ampicillin was added. Subsequently, 230 µL of methanol was added for protein precipitation and vortexed for 10 minutes. The sample was then centrifuged at 14,000 g for 10 minutes at 10°C. 100 µL supernatant was transferred to an Eppendorf vial and evaporated for 30 minutes in a Thermo-Speedvac (Thermo Scientific, San Jose, CA). The dried sample was resolved in 100 µL water containing 2 % acetonitrile and vortexed for 10 minutes.

Fragmentation tree acquisition, processing and database comparison
Fragmentation tree acquisition was carried out as previously described (**Chapter 5**) in positive ion mode on a LTQ-Orbitrap-XL mass spectrometer (Thermo Fisher Scientific, Waltham, MA) using an Advion Nanomate (Advion, Ithaca, NY) for nano-electrospray with following settings: 1,5kV source voltage, 120°C capillary temperature, 5V capillary voltage, 70V tube lens voltage and 35 % normalized collision energy for CID fragmentation. A minimum of three repetitions of the complete fragmentation tree were acquired.
Assigned fragmentation trees were created from acquired Thermo Xcalibur files using a macro programmed in Excel (Microsoft, Redmond, WA). Finally, all processed trees were compared to an in-house fungal metabolite MS^n database for determination of similar fragments and losses (Rojas-Cherto, et al., 2012).

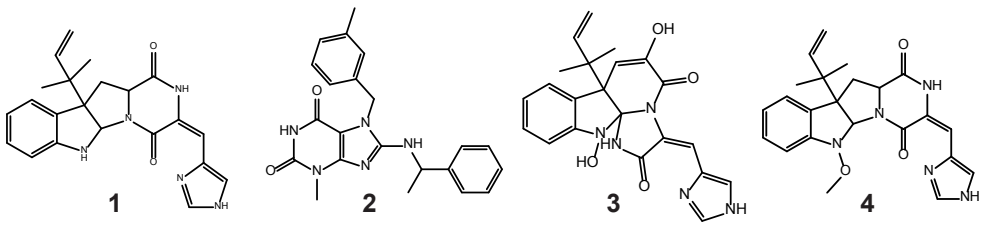


Figure 1. Structures of discussed compounds

Results and Discussion

To demonstrate the capability of CMCP to confidently elucidate the identity of unknown compounds in DI based experiments despite a co-fragmentation of interferences, the structure identification of the four analytes roquefortine C (**1**, $C_{22}H_{23}N_5O_2$, calc. $[M+H]^+$ = 390.1924), 3-me-7-(3-methylbenzyl)-8-((1-phenylethyl) amino)-3,7-dihydro-1H-purine-2,6-dione (**2**, $C_{22}H_{23}N_5O_2$, calc. $[M+H]^+$ = 390.1924) , glandicoline B (**3**, $C_{22}H_{21}N_5O_4$, calc. $[M+H]^+$ = 420.1666) and roquefortine F (**4**, $C_{23}H_{25}N_5O_3$, calc. $[M+H]^+$ = 420.2030) (Figure 1) is described under various conditions. First, structure identification based on identical database entries is shown for single, isomeric and isobaric ions, representing identity searches in DI experiments. To demonstrate *de novo* structure identification, in which the mass spectral data of the analyte is not represented in a database, the structure elucidation of **1** is shown using complementary fragment and neutral-loss tree comparison.

Fragmentation tree generation and database search

Multiple stage fragmentation spectra for each compound were acquired using direct infusion as described in the Materials and Methods section. Due to the gentle ionization of ESI, little fragmentation of precursor ions was observed. After elemental compositions were assigned to all fitting ions present in the acquired fragmentation tree using a fixed precursor composition, fragment and neutral-loss trees were automatically generated. The resulting trees were compared to an in-house MS^n database searching for similar mass fragments and neutral-losses. Obtained database entries were ranked in a descending order according their similarity represented by the Tanimoto coefficient (Flower, 1998).

Identity search of single ion trees

The database query of the fragment tree of compound **1**, which did not show additional abundant ions present in the same isolation range of MS^1 prior fragmentation, returned the database entry roquefortine C as most similar hit. As all fragments and neutral-losses of compound **1** could be found present and similarly arranged in the database entry, an identical structure for both compounds was indicated (Figure 2). This was supported by comparing their individual multiple stage fragmentation spectra showing almost identical mass spectra on every level (data not shown). Therefore, **1** was tentatively identified as roquefortine C which was ultimately confirmed using HPLC-MS/MS, comparing **1** and a standard of roquefortine C.

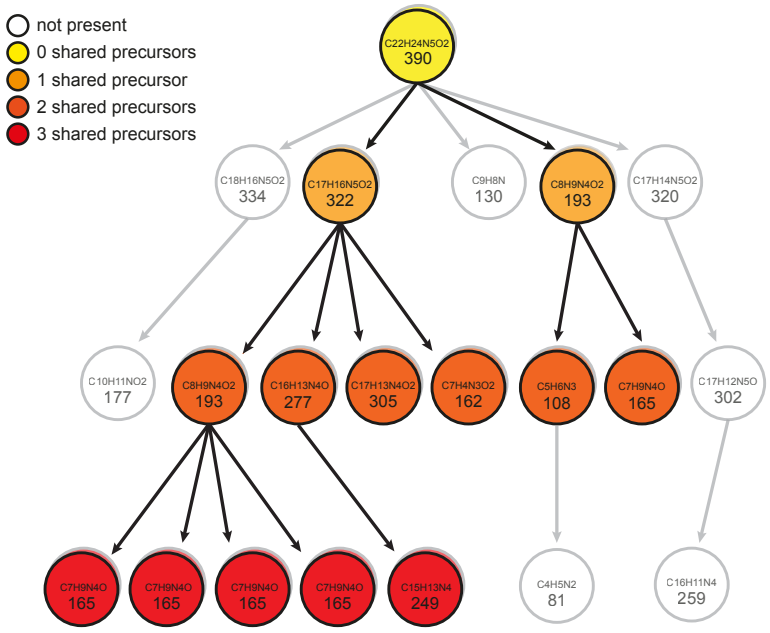


Figure 2. Fragment tree comparison without co-fragmented ions

Superimposed fragment trees of **1** (top), acquired from a biological samples using DI-MS, and partial fragmentation tree of database entry roquefortine C (back) after fragment tree comparison. All fragments of **1** could be found similarly arranged in the database entry (colored nodes) whereas not all fragments of the database entry could be found in the fragment tree of **1** (white nodes). Differences between both trees are due to different precursor intensities.

The difference in depth and width between the fragment tree of **1** and roquefortine C results from the different precursor intensities during fragmentation tree acquisition. The tree of **1** was acquired using a complex biological sample with a low concentration of **1**, whereas the fragmentation tree of roquefortine C in the reference database was acquired with a highly concentrated standard resulting in a much richer fragmentation.

Identity searches of isomeric ion trees

We next investigated the possibility of achieving unambiguous identification of co-fragmented isomeric precursor ions using CMCP by spiking the isomer **2** to a solution of **1**. With an identical mass-over-charge ratio in MS^1, the two isomers **1** and **2** with *m/z* 390.1926 ($C_{22}H_{23}N_5O_2$, calc. $[M+H]^+$ = 390.1924) are simultaneously fragmented resulting in one single fragmentation tree containing fragments from both compounds (Figure 3 and Figure 4A). Due to an identical elemental composition of their precursor ions in MS^1, all fragments originating from a fragmentation of **1** and **2** could be assigned with their correct elemental composition when processed with the chemical formula of protonated **1**. This resulted in the incorporation of all fragments of **2** into the fragment tree of **1** (Figure 4B). Using CMCP, the database query returned the protonated database entry roquefortine C ($C_{22}H_{23}N_5O_2$, $[M+H]^+$ = 390.1924) as most similar hit to **1**, sharing a common precursor ion next to several subtrees initiated in MS^2, thus indicating that **1** and roquefortine C are

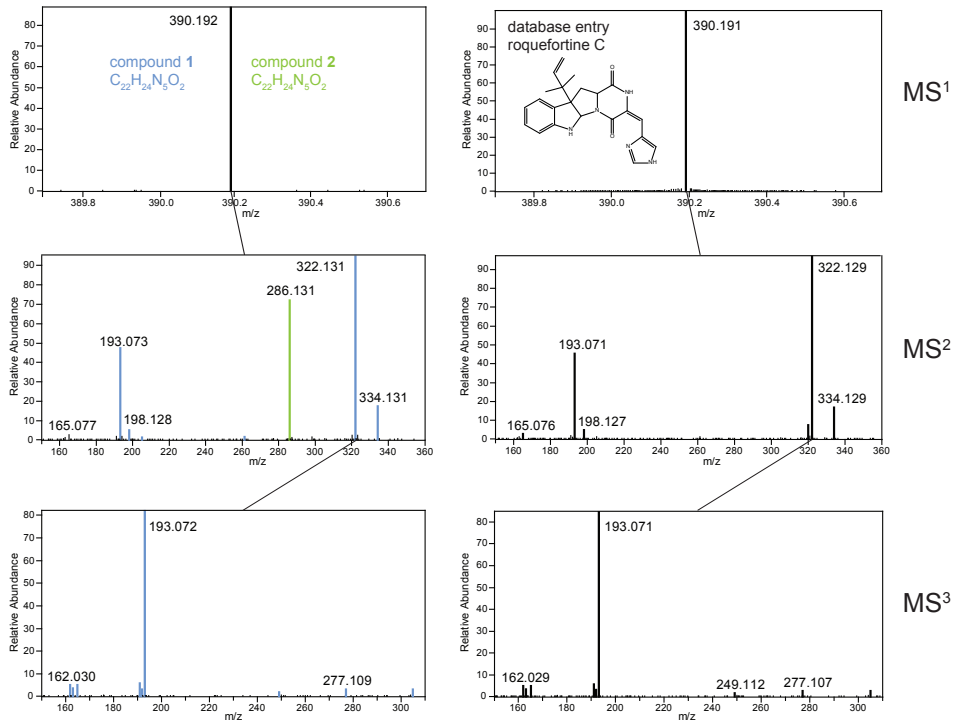


Figure 3. Multiple stage fragmentation spectra of the co-fragmented isomers **1** and **2** (left), and the database entry roquefortine C (right). Fragments originating from compound **1** are colored blue whereas fragments of **2** are colored green.

structurally related (Figure 4C). This was further supported by highly similar MS^2 spectra in which all fragments of roquefortine C were present with a similar mass-to-charge ratios and relative intensities in the MS^2 spectrum of co-fragmented **1** (Figure 3). In contrast, fragments originating from a co-fragmentation of **2**, like *m/z* 286.131 were highly abundant in the MS^2 of **1** but not present in the MS^2 spectra of the database entry roquefortine C, resulting in ambiguity about the structure of **1**. With the automatic multiple stage fragmentation spectra acquisition in CMCP, product ions in MS^2 were fragmented further yielding fragmentation spectra of various fragmentation stages, corresponding to a fragmentation of **1** which were free of interfering ions and completely identical to their analogue in the database entry roquefortine C. So shows the MS^3 spectra of *m/z* 322.131 in **1** and roquefortine C similar fragments in terms of mass-to-charge ratios and relative intensities (Figure 3, Figure 4C) leading to the conclusion that **1** and roquefortine C are structurally identical. In addition, fragments in the fragment tree of **1**, which were not found to be shared by the database entry roquefortine C, have to originate from a fragmentation of at least one additional isomer, namely compound **2** (Figure 4C).

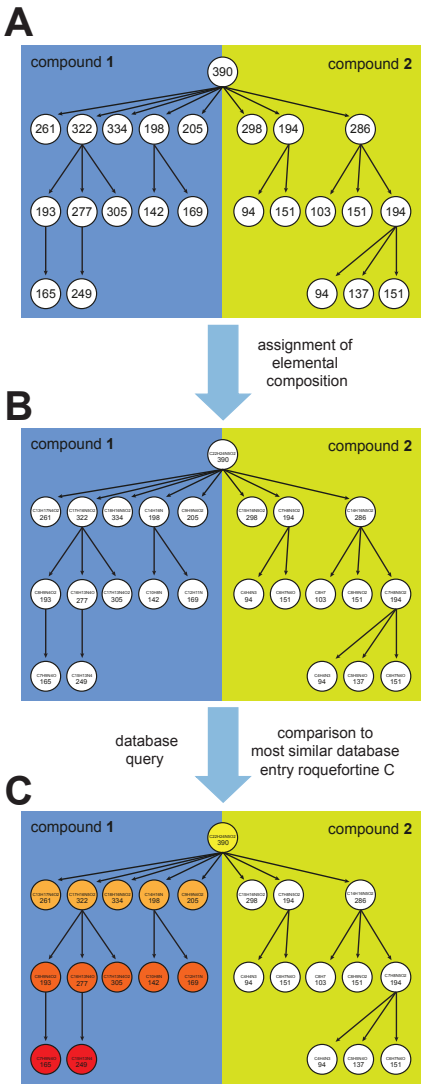


Figure 4. Structure elucidation of co-fragmented isomers. Fragment tree comparison for the co-fragmented isomers 1 and 2. As both compounds have the same elemental composition, all fragments of 2 were assigned with a correct elemental composition when processed with the precursor formula of protonated 1. The database query returned the database entry roquefortine C as most similar hit, with all fragments of 1 being present and similarly arranged.

In addition, as the complete spectral information of co-fragmented **2** is contained in the acquired fragmentation tree of **1**, the identification of **2** can be performed accordingly from the same multiple stage fragmentation data. Using the second most similar database entry 3-me-7-(3-methylbenzyl)-8-((1-phenylethyl) amino)-3,7-dihydro-1H-purine-2,6-dione from the database query, which shares solely the remaining fragments with the fragment tree of **1** and **2**, the structure of **2** could be determined.

These results demonstrate the confident identification of analytes in direct infusion using CMCP despite a co-fragmentation of isomeric ions.

Identity searches of isobaric ion trees

When analyzed separately, compound **3** with *m/z* 420.1659 ($C_{22}H_{21}N_5O_4$, cal. $[M+H]^+$ = 420.1666) could be unambiguously identified based on its fragmentation tree, following the strategy described for the Identity search of compound **1**. However, when analyzed in the liquid culture of *P. chrysogenum*, a compound with a very similar mass-over-charge ratio of 420.2024 (**4**, $C_{23}H_{25}N_5O_3$, calc. $[M+H]^+$ = 420.2030) was observed.

Although clearly distinguishable in MS^1, due to the high-resolution of the Orbitrap, compound **3** and the interfering isobaric ion of **4** were isolated together and fragmented simultaneously due to the insufficient resolving power available in the isolation step. This resulted in an identical fragmentation tree with identical MS^n spectra for both compounds, containing fragments originating from a fragmentation of **3** and **4** (Figure 5). By processing the acquired multiple stage fragmentation spectra with the chemical formula of protonated **3**, ions originating from a fragmentation of **3** were assigned with their correct elemental composition and incorporated into the fragmentation tree. In contrast, fragments originating from a fragmentation of **4** were mainly rejected as their elemental composition was not consistent with the chemical formula of **3** (Figure 6A and B). So is the fragment with *m/z* 389 ($C_{22}H_{23}N_5O_2$) of **4** and its complete subtree, not present in the fragmentation tree of **3** as its corresponding elemental composition contains two additional hydrogens com-

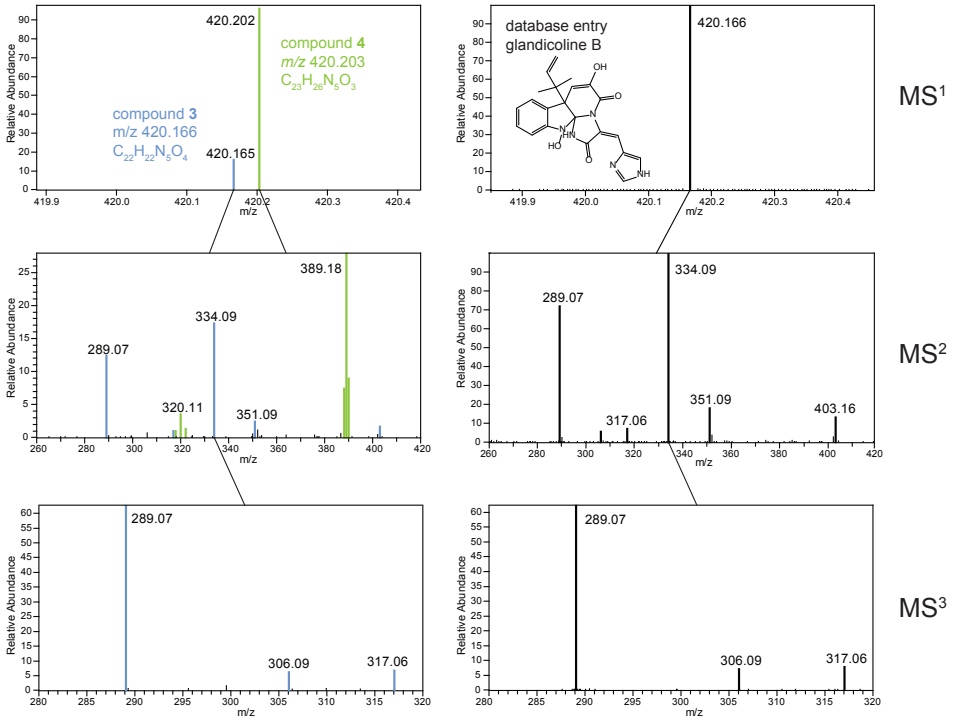


Figure 5. Multiple stage fragmentation spectra of the co-fragmented isobars **3** and **4**, acquired from a biological sample using DI-MS (left), and the database entry glandicoline B (right). Fragments originating from compound **3** are colored blue whereas fragments of **4** are colored green.

pared to protonated **3**. However, due to the relatively similar elemental composition of **3** and **4**, few fragments originating from a fragmentation of **4** could be still assigned, like *m/z* 320 ($C_{17}H_{14}N_5O_2$), *m/z* 350 ($C_{18}H_{16}N_5O_3$) and *m/z* 388 ($C_{22}H_{22}N_5O_2$). The subsequent database query of **3** returned highest similarity to the database entry glandicoline B ($C_{22}H_{21}N_5O_4$, $[M+H]^+$ = 420.16663) which shares next to similarly arranged fragments and neutral-losses also its precursor ion with **3** (Figure 5 and Figure 6C). MS^2 fragments, present in both compounds, showed similar subtrees, indicating that **3** and glandicoline B share a similar (sub)structure. This was supported by comparing their individual multiple stage fragmentation spectra showing almost identical relative intensities of fragments shared by **3** and glandicoline B (Figure 5). Therefore, it is concluded that remaining fragments in the fragment tree of **3**, which were not found shared by the database entry glandicoline B, correspond to a fragmentation of **4** (Figure 6C). The structure of **3** was tentatively identified as glandicoline B and ultimately confirmed using HPLC-MS/MS comparing **3** and a chemical standard of glandicoline B.

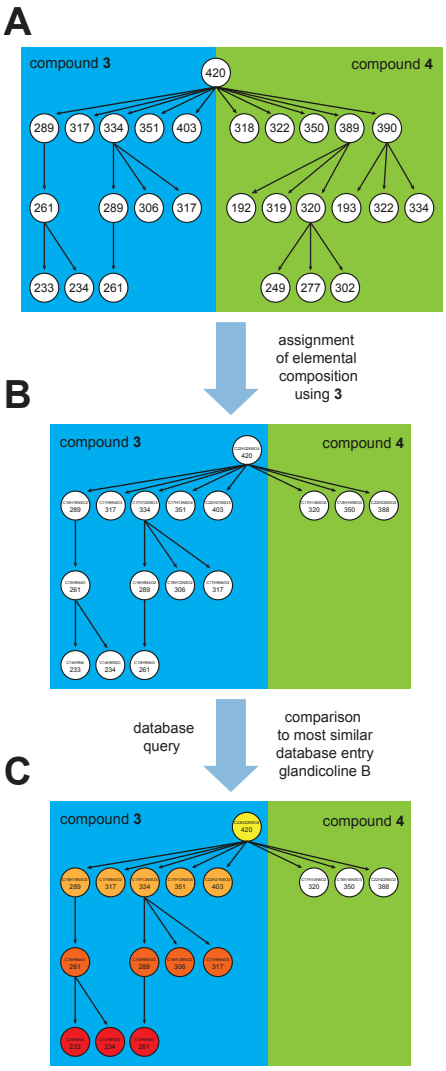


Figure 6. Structure elucidation of co-fragmented isobars
Fragment tree comparison of the co-fragmented isobars **3** and **4**. First, an unassigned fragmentation tree containing fragments of compound **3** and **4** was obtained. During the assignment of elemental compositions to fragments using the precursor formula of **3**, several fragments originating from **4** were rejected due to inconsistency. Last, the database query returned the database entry glandicoline B as most similar entry with all fragments of **3** being present and similarly arranged. The complete fragment trees of **3** and **4** can be found in the Supplement.

As **3** and **4** were fragmented simultaneously, the structural information of **4** is contained in the same multiple stage fragmentation spectra which were used to identify **3**. By processing the fragmentation tree of co-fragmented **3** and **4** with the elemental composition of **4**, instead of **3**, the fragment and neutral-loss tree of **4** were obtained (data not shown). Their comparison to the MS^n database returned roquefortine F as most similar database entry sharing almost all fragments and neutral-losses with **4**. Ultimately, **4** was tentatively identified as roquefortine F which was confirmed using NMR.

The structure identification of co-fragmented isobaric ions is comparable to the identification of compounds from single or isomeric ions, depending on

the elemental composition and fragmentation of the compounds. In case none of the fragments in MS^2 of the interfering isobaric compound are consistent with the precursor formula of the targeted analyte, a fragment tree without interfering fragments is obtained for the analyte resembling the identification of a single fragmented compound. If however, both isobars share a similar elemental composition, fragments and neutral-losses of the co-fragmented isobaric ions can be incorporated into the processed trees of the analyte, resembling the fragment and neutral-loss trees of co-fragmented isomers. Therefore, similar strategies for their identification apply.
In contrast to the identification of one population of structurally identical ions or co-fragmenting isomeric ions, the presence of isobaric ions is immediately recognized as such, as more than one precursor ion in MS^1 can be observed if the resolution of the mass spectrometer is sufficient enough.

De novo structure identification
So far, the structure elucidation shown was achieved using an identical standard present in the MS^n database. However, an identical database entry is not always available, especially when completely new compounds with novel structures were discovered. To demonstrate the *de novo* structure identification of single fragmented **1**, using not an identical but similar database entry, the database entry of **1** (roquefortine C) was removed from the database. Without an identical entry present in the database, the database query of **1** with *m/z* 390 ($C_{22}H_{23}N_5O_2$, calc. $[M+H]^+ = 390.1924$), returned highest similarity to the database entry dehydrohistidyltryptophanyldiketopiperazine (DHTD) with *m/z* 322 ($C_{17}H_{15}N_5O_2$, calc. $[M+H]^+ = 322.1299$), which shares almost its complete fragment and neutral-loss trees with **1**, as well as the database entry roquefortine D with *m/z* 392 ($C_{22}H_{25}N_5O_2$, calc. $[M+H]^+ = 392.2081$) which shares primarily a consecutive neutral-loss path (Figure 7A and 7B). As both database entries have a different elemental composition as **1**, it can be immediately concluded that a similar substructure rather than an identical structure is shared. Based on their shared fragments and neutral-losses, similar fragmentation mechanisms could be identified and a (sub)structure of **1** deduced (**Chapter 5**). As all fragments and neutral-losses in the fragment and neutral-loss tree of **1** originate from a fragmentation of **1**, the remaining unshared fragments can be used to validate the deduced structure for consistency, providing additional confidence for the proposed structure.
Different to the fragmentation tree of single fragmented **1**, in which all fragments originate from a fragmentation of **1**, the fragmentation tree of the co-fragmented isomers **1** and **2** contains additional fragments and neutral-losses from a fragmentation of **2** (Figure 7A and 7B). Although the same fragments and neutral-losses are shared with the database entries DHTD and roquefortine C, as shown for **1** without a co-fragmented interfering ion, a lower similarity value is obtained due to the additional unshared fragments and neutral-losses of **2**. Based on the shared nodes and the known structure of the database entry, a common fragmentation mechanism could be identified and a (sub)structure deduced for **1** (**Chapter 5**). In contrast to the previous identification, not all unshared fragments and neutral-losses can now be used to support the tentatively identified structure of **1** as they might originate from a fragmentation of **2**. However, a complete inconsistency with the tentatively

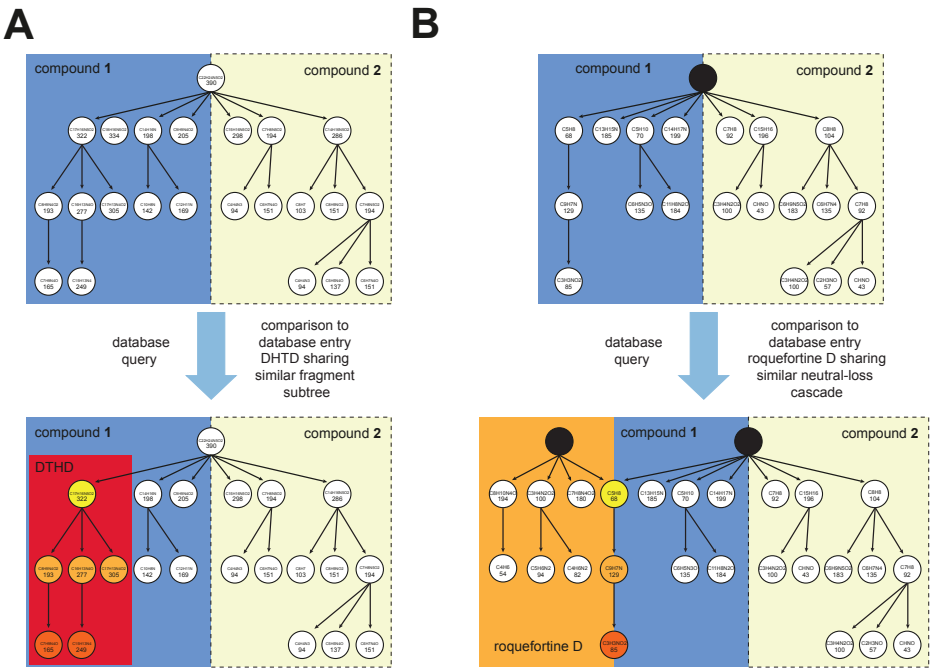


Figure 7. Schematic *de novo* identification of compound **1**, co-fragmented with the isomer **2**, using fragment tree comparison to the database entry DHTD (A) and neutral-loss tree comparison to the database entry roquefortine D (B). By removing the fragments and neutral-losses of **2**, indicated by a dashed square, the *de novo* identification of **1** without a co-fragmented ion can be deduced.

identified structure of **1** can give an indication that structurally different fragments are present most likely originating from a co-fragmentation of an isomeric precursor ion.

Comparable to the identity search of co-fragmented isobaric ions, shows the *de novo* structure elucidation of co-fragmented isobars either more similarity to the *de novo* identification of compounds from single ions or isomeric ions, depending on the elemental composition and fragmentation of the involved compounds. Although the presence of isobaric ions can be clearly recognized based on the presence of more than one precursor ion in MS^1, it can pose a similar challenge as the *de novo* identification of co-fragmented isomers if all fragments of the interfering isobar can be assigned with a valid elemental composition when processing with the molecular formula of the analyte. Therefore, similar strategies for their identification apply.

Conclusions

As a preceding separation step is absent in DI, all ions originating from a particular sample are, in general, measured at the same time in the mass spectrometer. Depending on the composition of the sample, unwanted ions can interfere with the compound of interest in form of isobaric or isomeric ions, resulting in their simultaneous fragmentation hampering structural identification. Different to MS/MS

approaches, fragment ions of an analyte can be structurally characterized using the Chemoinformatics supported MS^n Comparison Pipeline (CMCP) due to consecutive isolation and fragmentation allowing their differentiation from ions originating from co-fragmenting isobaric or isomeric precursor ions. With the subsequent generation of fragment and neutral loss trees from the acquired multiple stage fragmentation spectra, sub-trees from co-fragmented isobars can be confidently removed resulting in fragmentation data solely originating from the analyte. In cases where the sample compound is present in the database, a fully automated identification is possible despite co-fragmented interferences enabling detailed structure elucidation in direct-infusion based experiments. This makes CMCP a valuable tool as also newer generation mass spectrometer will not be able to completely avoid a co-fragmentation of isomeric and isobaric ions.
Furthermore, as CMCP is searching for common fragments and neutral-losses, also database entries which don't fragment in an identical, but similar fashion are returned. With the structural elucidation of shared fragments, deduced from the known structure of the database entry, confident *de novo* identification of unknown molecular (sub)structures can be achieved, even without the compound being present in the database. Therefore, less comprehensive databases are required to cover a large chemical space as not for every analyte an identical standard needs to be present in the database.
With the facile integration into already existing profiling pipelines, structural information can be directly obtained from DI experiments, offering more than just a 'first pass' screening without the immanent need to switch to another platform.

Acknowledgements

This project was supported by the Perspective Genbiotics program subsidized by the Stichting voor de Technische Wetenschappen (STW) and (co)financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

References

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K., et al. (2012). KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol 53, e1.

Boernsen, K.O., Gatzek, S., and Imbert, G. (2005). Controlled protein precipitation in combination with chip-based nanospray infusion mass spectrometry. An approach for metabolomics profiling of plasma. Analytical chemistry 77, 7255-7264.

Cui, M., Sun, W., Song, F., Liu, Z., and Liu, S. (1999). Multi-stage mass spectrometric studies of triterpenoid saponins in crude extracts from Acanthopanax senticosus harms. Rapid communications in mass spectrometry : RCM 13, 873-879.

Dettmer, K., Aronov, P.A., and Hammock, B.D. (2007). Mass spectrometry-based metabolomics. Mass Spectrom Rev 26, 51-78.

Draper, J., Lloyd, A.J., Goodacre, R., and Beckmann, M. (2013). Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: a review. Metabolomics 9, 4-29.

Flower, D.R. (1998). On the Properties of Bit String-Based Measures of Chemical Similarity. J. Chem. Inf. Comput. Sci 38, 379-386.

Goodacre, R., Vaidyanathan, S., Bianchi, G., and Kell, D.B. (2002). Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. Analyst 127, 1457-1462.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40, D109-114.

Kang, J., Hick, L.A., and Price, W.E. (2007). A fragmentation study of isoflavones in negative electrospray ionization by MSn ion trap mass spectrometry and triple quadrupole mass spectrometry. Rapid Commun Mass Spectrom 21, 857-868.

Kasper, P.T., Rojas-Cherto, M., Mistrik, R., Reijmers, T., Hankemeier, T., and Vreeken, R.J. (2012). Fragmentation trees for the structural characterisation of metabolites. Rapid Commun Mass Spectrom 26, 2275-2286.

Koulman, A., Tapper, B.A., Fraser, K., Cao, M., Lane, G.A., and Rasmussen, S. (2007). High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. Rapid Commun Mass Spectrom 21, 421-428.

Li, Q., Cheng, T., Wang, Y., and Bryant, S.H. (2010). PubChem as a public resource for drug discovery. Drug Discov Today 15, 1052-1057.

Mungur, R., Glass, A.D., Goodenow, D.B., and Lightfoot, D.A. (2005). Metabolite fingerprinting in transgenic Nicotiana tabacum altered by the Escherichia coli glutamate dehydrogenase gene. J Biomed Biotechnol 2005, 198-214.

Nakamura, Y., Kimura, A., Saga, H., Oikawa, A., Shinbo, Y., Kai, K., Sakurai, N., Suzuki, H., Kitayama, M., Shibata, D., et al. (2007). Differential metabolomics unraveling light/dark regulation of metabolic activities in Arabidopsis cell culture. Planta 227, 57-66.

Raterink, R.J., Lindenburg, P.W., Vreeken, R.J., and Hankemeier, T. (2013). Three-Phase Electroextraction: A New (Online) Sample Purification and Enrichment Method for Bioanalysis. Analytical chemistry.

Rochfort, S.J., Trenerry, V.C., Imsic, M., Panozzo, J., and Jones, R. (2008). Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. Phytochemistry 69, 1671-1679.

Rojas-Cherto, M., Peironcely, J.E., Kasper, P.T., van der Hooft, J.J., de Vos, R.C., Vreeken, R., Hankemeier, T., and Reijmers, T. (2012). Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. Analytical chemistry 84, 5524-5534.

Savitski, M.M., Sweetman, G., Askenazi, M., Marto, J.A., Lang, M., Zinn, N., and Bantscheff, M. (2011). Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers. Analytical chemistry 83, 8959-8967.

Wichitnithad, W., McManus, T.J., and Callery, P.S. (2010). Identification of isobaric product ions in electrospray ionization mass spectra of fentanyl using multistage mass spectrometry and deuterium labeling. Rapid Commun Mass Spectrom 24, 2547-2553.

Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., et al. (2009). HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37, D603-610.

Chapter 7

Summary, conclusions and perspectives
Nederlandse samenvatting

Summary, conclusions and perspectives

With the persistent endeavor to provide faster, healthier, cheaper and more targeted drugs, there is a constant need for novel bioactive compounds with new chemical scaffolds. Although already exploited for more than a century, plants and microorganisms are still assumed as a major source of novel and innovative therapeutic agents with antimicrobial, anticancer, antifungal or immunosuppressive properties (Newman and Cragg, 2007). However, classical bio-assay guided drug discovery strategies, which were used for the last 70 years, are generally long, cost intensive and laborious, with a high rediscovery rate of known natural products, which resulted in a mass-withdrawal of pharmaceutical companies from natural product discovery over the last years (Zerikly and Challis, 2009). With the substantial cost reduction and advances in DNA sequencing, a huge quantity of DNA sequence data from a wide variety of organisms is now available, offering novel strategies and techniques for the discovery of potentially novel bioactive compounds. Genome mining, which screens for genetic sequences possibly encoding enzymes that are likely to be involved in the biosynthesis of bioactive natural compounds represents such a novel and powerful approach (Zerikly and Challis, 2009). Its combination, for instance with heterologous gene expression and comparative metabolite profiling have already proven successful for the discovery, isolation and characterization of novel bioactive metabolites (Bok, et al., 2006). Although the development of structural analysis and purification techniques has significantly accelerated the process of natural product discovery, metabolite discovery and especially metabolite identification pose still a major challenge in natural product discovery requiring novel techniques and analytical platforms.
In this thesis the development of an analytical pipeline for the discovery, identification and extraction of novel natural products from fungal fermentation cultures is described. With structure elucidation being the major bottleneck, the focus was set on the implementation and further development of innovative methods for the complete, fast and sensitive identification of complex metabolites. These developed analytical platforms were successfully applied for the discovery, the identification and ultimately the assignment of secondary metabolites to their corresponding NRPS genes and gene clusters in the fungus Penicillium chrysogenum.

In **Chapter 2**, the identification of natural products from a di-modular non-ribosomal peptide synthetase (NRPS) cluster in Penicillium chrysogenum and their corresponding enzymatic steps are described. With five additional tailoring enzymes that modify the initial NPRS product, structures of natural products could not be deduced directly from the genetic sequence. Thus, individual deletion of the naturally expressed biosynthetic genes was performed followed by comparative metabolite profiling. In order to discover and identify the unknown natural products originating from this NRPS cluster, an unbiased, sensitive and fast analytical method was required. Using reversed phased liquid chromatography (RP-LC) in combination with high resolution and high accuracy mass spectrometry, obtained by a linear ion-trap Fourier transform mass spectrometer (LTQ-FTMS), a untargeted metabolite profiling method was developed for the discovery of medium-polar to unpolar compounds. Based on a full analytical validation, it was concluded that the method

was suitable for the detection of novel natural products with good reproducibility, linearity, high coverage and high sensitivity enabling metabolite discovery at nanomolar levels. This method was therefore used throughout this thesis, enabling the discovery of about 40 relevant secondary metabolites from two different NRPS clusters in Penicillium chrysogenum (**Chapter 2** – **4**).
Subsequent LC-MSn and NMR analysis of the discovered metabolites originating from the di-modular NRPS revealed astonishing compounds with complex chemical scaffolds, harboring among others antimicrobial properties. By linking their production in host and derived deletion strains to the performed genetic modification, a first step towards the identification of their biosynthetic machinery could be achieved.

In **Chapter 3**, the structural identification of five lower concentrated products obtained from comparative metabolite profiling of deletion strains from the di-modular NRPS in **Chapter 2** is described. Based on their novel structures missing links in the biosynthetic machinery could be filled, revealing that all involved enzymes of this pathway are rather unspecific by catalyzing more than one reaction. This results in natural products with unusual chemical structures harboring cytotoxic and potentially anti-cancer properties. Starting with histidyltryptophanyldiketopiperazine (HTD), synthesized by the core synthetase enzyme RoqA using tryptophan and histidine as substrates, RoqD catalyzes the biosynthesis of roquefortine D from HTD. At the same time, RoqR, a cytochrome p450 oxidoreductase, oxidizes HTD at its histidinyl moiety to dehydrohistidyltryptophanyldiketopiperazine (DHTD). Both simultaneous reactions of HTD lead to a branch of the roquefortine/meleagrin pathway, one to DHTD via the oxidation by RoqR and further to roquefortine C by dimethylallyl addition of RoqD, and the other via an alteration of the enzymatic order. There, dimethylallyl addition is first performed by RoqD to yield roquefortine D while further oxidation is carried out by RoqR yielding roquefortine C. RoqM, encoding a flavin-dependent MAK 1-monooxygenase like protein, was found to be involved in the conversion of roquefortine C to roquefortine L, a novel compound containing an unusual nitrone moiety, possibly with a hydroxyroquefortine like structure as intermediate. Additionally, Roquefortine F, previously only reported from a deep ocean sediment derived Penicillium species, is synthesized from a consecutive enzymatic modification of RoqM and RoqN on roquefortine C, yielding ultimately neoxaline involving RoqO. In addition, RoqN and RoqO were also found be involved in the biosynthesis of glandicoline B and meleagrin, however, acting in a reversed order. In summary, these results demonstrated the suitability of the developed pipeline for the discovery and de novo identification of novel metabolites from complex biological samples.

Next, the metabolite discovery and identification pipeline was successfully applied for the investigation of a cryptic non-linear tetra-modular NRPS present in Penicillium chrysogenum, which is described in **Chapter 4**. By deleting the putative gene in combination with comparative metabolite profiling various novel cyclic and linear tetrapeptides with similar chemical properties were identified and subsequently associated to this NRPS. Based on the sequence of the produced cyclic products, distinctive unspecificity of all involved adenylation domains towards their respec-

tive substrates was deduced resulting in an incorporation of different amino acids into the final product. In combination with substrate predictions for each module, the detailed mechanism for their production could be determined. This study demonstrates the need for comprehensive analytical methods which allow profound insight into the metabolome. As the biosynthetic mechanism was deduced from the amino acid sequence of the products, comprehensive quantitative and structural information of the natural products was required to link each adenylation domain correctly to its individual chain extending step.

However, de novo structure elucidation as applied in **Chapter 2** - **4** is a challenging task which either offers only limited structural information, as often observed for mass spectrometry despite MS/MS fragmentation, or demands high sample purity and amounts as required for NMR. To overcome these limitations, first steps towards an automated de novo structure elucidation were taken. In **Chapter 5**, the structure elucidation pipeline CMCP (Chemoinformatics supported MS^n Comparison Pipeline), is presented which is based on the comparison of multiple stage fragmentation mass spectrometry (MS^n) data. After metabolite extraction and high-resolution fragmentation tree acquisition, fragment and neutral-loss trees of unknowns are generated and compared to a MS^n database using computational tools to extract compounds with similar fragmentation behavior, represented by shared fragments and/or neutral-losses. Identifying these shared fragmentation mechanisms, based on the known structure of the database entry, allows to confidently deduce chemical (sub)structures of unknowns, even without the need of the unknown compound to be present in the database. Depending on the results from the database query of the unknown compound, similarity can be a result of identical or similar fragments, identical or similar neutral-losses or a combination of both. To demonstrate the different conceptual outcomes of a database query (similarity based on an identical fragment and neutral-loss tree, similarity based on a similar fragment and neutral-loss tree, similarity based on a similar neutral-loss tree) and to show how this information can be used to deduce structural information, the de novo identification of the previously unknown secondary metabolites roquefortine C, dehydrohistidyltryptophanyldiketopiperazine (DHTD), roquefortine F and roquefortine D, obtained from comparative metabolites profiling, was shown. Their tentatively identified (sub)structures were ultimately confirmed by NMR.

Traditional approaches for the discovery of metabolites by mass spectrometry use chromatographic separation like LC (Liquid Chromatography) before metabolite detection, as described in **Chapter 2** - **4**. Due to the chromatographic separation step, laborious and time consuming sample preparation procedures are required next to significant analysis times. Furthermore, separation and detection of metabolites is heavily depending on the chemical nature of the analytes and the used chromatographic method. In contrast, direct infusion mass spectrometry (DI-MS) is a fast, reliable, sensitive and cost effective method which is particularly attractive when dealing with large sample sets. However, structure identification in direct infusion based experiments is still a major obstacle due to co-fragmenting interferences like isobaric and isomeric ions, limiting most applications to metabolite fingerprinting only. In **Chapter 6**, the further development and application of CMCP for the struc-

ture elucidation of metabolites in direct infusion based experiments was described. Similar to the structure elucidation demonstrated in **Chapter 5**, molecular structures of unknowns are deduced from shared fragmentation mechanisms of similar fragmenting database entries. However, in contrast to the structure elucidation presented in **Chapter 5**, in which pure sample fractions were used for fragmentation tree acquisition, in direct infusion unwanted ions can interfere with the compound of interest in form of isobaric or isomeric ions due to a missing preceding chromatographic separation step. As they are fragmented simultaneously, MS^2 fragmentation spectra with additional fragments are obtained which hamper structural identification. However, by fragmenting these fragments further, unwanted fragments of interfering ions are removed via successive isolation and fragmentation, resulting in subtrees originating ideally from only one population of structurally identical ions, as shown for the identification of roquefortine C, roquefortine F and glandicoline B. With CMCP searching for common fragments and neutral-losses over various fragmentation stages, database entries with similar subtrees were extracted, which allowed the structural identification of (sub)structures of the unknown compound, even without the compound being present in the database. With the facile integration of a DI profiling approach, metabolite discovery and identification can be combined into one analytical platform, extending DI to more than just a 'first pass' screening without the immanent need to switch to another platform.

In conclusion, this thesis demonstrated the development and application of analytical platforms for the discovery, relative quantification and identification of secondary metabolites from Penicillium chrysogenum. Although solely applied to fungal liquid cultures, these platforms can be easily extended to meet the requirements for the discovery and identification of metabolites from other biological samples.
Especially the utilization of automated fragmentation tree comparison has proven to be very powerful for the structural identification of metabolites. Compared to traditionally used MS/MS approaches in which product ions are acquired into one spectrum, MS^n approaches fragment product ions further giving insight into precursor-product ion relationships. As the structure of a precursor ion is coded in its fragments, subtrees contain structural information of their precursors. Thus, the comparison of fragments in MS^n is not limited to mass-over-charge ratios, elemental compositions or intensities, but in addition, allows the comparison of fragment structures giving a tremendous advantage compared to MS/MS approaches. Besides being clearly beneficial for *de novo* identifications, as shown in **Chapter 5** and **6**, multiple stage fragmentation tree comparison also offers much more confidence in traditional identity searches, in which the database entry is present in the database, as fragments and neutral-losses are compared over multiple fragmentation stages rather than only on one level. Therefore, it is expected that MS^n methods will more and more replace MS/MS based approaches in the near future in cases where detailed structural information of analytes is needed but only limitedly available. However, due to much longer acquisition times of MS^n experiments compared to MS/MS approaches, complete fragmentation trees can't be always generated on the fly during metabolite profiling with the current generation of mass spectrometers. Therefore, methods like on the fly fraction collection followed by direct infusion mass spectrometry, as offered by the Advion Nanomate, direct infusion

mass spectrometry without prior separation, as described in **Chapter 6** or repetitive HPLC-MS^n experiments in which specific subtrees of a fragmentation tree are acquired during each run which are ultimately assembled to a complete tree, can be combined with existing profiling approaches to efficiently integrate multiple stage fragmentation tree comparison into profiling platforms.

However, it should be pointed out that comparison of fragmentation trees is still a time consuming task which requires profound chemical knowledge of fragmentation mechanisms. If not a completely identical but a merely similar database entry was obtained from fragmentation tree comparison, individual fragmentation mechanisms, shared by both compounds, had to be identified based on the known structure of the database entry, as described in **Chapter 5**. Unfortunately, the knowledge of fragmentation mechanisms obtained by chemical induced dissociation (CID), which was used to enable multiple stage fragmentation, is still limited and cannot be exactly predicted with algorithms, although software packages and algorithms are available to predict to some extent fragmentation. Different to fragmentation obtained from electron ionization (EI, formerly known as electron impact) which is extensively described in literature, the identification of fragments obtained by CID is still often done manually by experts. However, with larger MS^2 and MS^n databases becoming available, it is expected that mining these databases can lead to a deeper understanding of collision induced dissociation mechanisms, which can be translated into exact fragmentation predictions, ultimately accelerating mass spectrometry based structure elucidation. An alternate approach is the structural elucidation of fragments and neutral-losses in database entries by MS experts. By storing detailed structures of nodes into the MS^n database, structures of similar nodes present in unknown target compounds can be automatically assignment in later database queries, enabling an immediate identification of fragments and neutral-losses without further intervention of MS experts. By automatically combining the structural overlap of various subtrees or neutral-losses from one or more similar database entries, complete (sub)structures could be deduced allowing a completely automated structure identification. Thus, multiple stage fragmentation tree comparison is an extremely valuable identification approach which is expected to have a substantial impact on metabolite identification. Especially in natural product drug discovery, which aims to find novel natural products with completely new chemical scaffolds, which are not present in any database, multiple stage fragmentation tree comparison offers a fast, sensitive, cost effective and reliable solution.

With combinatorial chemistry failed to deliver leads that form the basis for development of successful new drugs, other strategies need to be found for drug discovery. With the substantial cost reduction and advances in DNA sequencing, a huge quantity of DNA sequence data from a wide variety of organisms became available opening new possibilities in natural product drug discovery, like shown in **Chapter 2** - **4**. As the tremendous increase in genome mining based studies over the last years already indicates, a resurgence of interest in natural product drug discovery is expected. However, the majority of these studies is very suggestive and confined as their metabolite discovery step is limited due to the used analytical methods. Especially the prevalent application of spectroscopy based detection methods like UV-

Vis heavily reduces the amount of detectable metabolites, as only a limited number of analytes absorb light within the UV-Vis range efficiently. In addition, commonly targeted methods are used for metabolite discovery which allow solely a detection of selected metabolites. Novel products or intermediates obtained from genetic modifications are, however, mostly completely neglected.
Therefore, different analytical tools are necessary to obtain comprehensive chemical information which requires experts from different analytical fields like metabolite profiling (on GC, LC, CE coupled to UV-Vis, MS etc.), metabolite identification (MS^n, NMR, etc.), statistical analysis and chemoinformatics working closely together in a state of the art environment. Next to a high level of expertise, efficient, coordinated and structured processes are required to ultimately answer urgent biological, pharmacological or medicinal questions using the appropriate analytical technology and strategies.
However, extensive cooperation is not only limited to the analytical part. As genome-based metabolite discovery combines multiple disciplinary processes like genome mining, genetic modification, comparative metabolite profiling, metabolite identification, pathway construction, bioactivity testing etc., scientific experts from each discipline are required to efficiently cooperate to unravel the mechanisms of secondary metabolite formation and the bioactivity of intermediates and products produced. It is expected that such collaborations will allow to discover and identify novel bioactive natural products for the development of new medicines.

References

Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D., and Keller, N.P. (2006). Genomic mining for Aspergillus natural products. Chemistry & biology 13, 31-37.

Newman, D.J., and Cragg, G.M. (2007). Natural products as sources of new drugs over the last 25 years. J Nat Prod 70, 461-477.

Zerikly, M., and Challis, G.L. (2009). Strategies for the discovery of new natural products by genome mining. Chembiochem 10, 625-633.

Nederlandse Samenvatting

Met het aanhoudende streven naar snellere, gezondere, goedkopere en meer gerichte geneesmiddelen, is het noodzakelijk en erg belangrijk om nieuwe bioactieve verbindingen en chemische samenstellingen te identificeren. Planten en micro-organismen worden al meer dan een eeuw gebruikt als één van de belangrijkste bronnen voor nieuwe en innovatieve therapeutische middelen met een anti-kanker, antischimmel of immuun onderdrukkende werking (Newman and Cragg, 2007). Echter de klassieke bio-assays welke de laatste 70 jaar zijn toegepast voor het ontdekken van gerichte geneesmiddelen zijn tijdrovend, kostbaar en erg omslachtig. Verder is de kans erg groot dat een al eerder bekend natuurlijk product wordt herontdekt. Mede hierdoor is de laatste jaren een afname in de interesse van farmaceutische bedrijven geconstateerd voor de ontdekking en ontwikkeling van natuurlijke producten (Zerikly and Challis, 2009). Door de aanzienlijke kostenbesparing en vooruitgang in de DNA sequentietechnieken is er een grote hoeveelheid DNA sequentiegegevens voor een verscheidenheid aan organismen beschikbaar gekomen waardoor er nieuwe strategieën en technieken zijn voor de ontdekking van nieuwe bioactieve verbindingen. Het zogenoemde “genome mining” zorgt voor de screening van genetische sequenties om zo mogelijke enzymen te identificeren welke betrokken zijn bij de biosynthese van bioactieve natuurlijke verbindingen, wat zorgt voor een krachtige aanpak (Zerikly and Challis, 2009). De combinatie met bijvoorbeeld heterologe genexpressies en vergelijkende metabolietprofielen heeft al bewezen succesvol te zijn voor de ontdekking, isolatie en karakterisering van nieuwe bioactieve metabolieten (Bok et al., 2006). De ontwikkeling van structuuranalysetechnieken en zuiveringstechnieken hebben een significante sprong genomen in het proces van de natuurlijke product ontdekking, metaboliet ontdekking en vooral in de metaboliet identificering maar vormen nog steeds een belangrijke uitdaging in de ontdekking van natuurlijke producten waarvoor nieuwe technieken en analytische methoden nodig zijn.

In dit proefschrift beschrijven we de ontwikkeling van een analytische “pipeline” voor de ontdekking, isolatie en identificatie van nieuwe natuurlijke producten uit schimmelfermentatie-culturen. Omdat de opheldering van moleculaire structuren altijd een belangrijk knelpunt is geweest, lag de focus van dit onderzoek voornamelijk op de ontwikkeling en toepassing van innovatieve methoden voor complete, snelle en gevoelige identificatie van complexe metabolieten. De ontwikkelde analytische platforms werden met succes toegepast bij de ontdekking en identificatie van secundaire metabolieten. Hierdoor konden deze structuren uiteindelijk geassocieerd worden aan de gerelateerde di-modular non-ribosomal peptide synthetase (NRPS) genen en genenclusters in de schimmel Penicillium chrysogenum.

In **Hoofdstuk 2** beschrijven we de identificatie van natuurlijke producten van een di-modulair non-ribosomaal peptide synthetase (NRPS) cluster en de daarbij behorende enzymatische omzettingen in *Penicillium chrysogenum*. Aangezien het niet mogelijk was om, gebruik makend van vijf specifieke enzymen die de originele NRPS producten modificeren, de structuren van deze verbindingen te herleiden en deze te associeëren aan de genetische informatie, is een vergelijkende metaboliet identi-

ficatie studie uitgevoerd in schimmelculturen waar individuele genen, verantwoordelijk voor de biosynthese, werden verwijderd of ge-inactieveerd. Om onbekende natuurlijke producten, afkomstig van deze NRPS clusters, te herkennen en te identificeren is een gevoelige en snelle analytische methode vereist. Door gebruik te maken van "omgekeerde fase" vloeistof chromatografie (RP-LC) in combinatie met een hoge resolutie en hoge nauwkeurigheid massaspectrometer (linear ion-trap Fourier transform mass spectrometer (LTQ-FTMS)) werd er een metaboliet profileringsmethode ontwikkeld om medium-polaire en apolaire stoffen op te sporen. Na volledige analytische validatie concluderen we dat deze methode, met een goede reproduceerbaarheid, lineariteit en gevoeligheid, geschikt is voor de detectie van nieuwe natuurlijke producten waarmee een breed spectrum aan metabolieten op het nanomolair concentratie niveau aantoonbaar is. Deze methode is, zoals omschreven in dit proefschrift, verder gebruikt bij het onderzoek om circa 40 relevante secundaire metabolieten te ontdekken en te identificeren in twee verschillende NRPS clusters in *Penicillium chrysogenum* (**Hoofdstuk 2-4**).
De gevonden metabolieten, afkomstig van het di-modulaire NRPS, werden later geanalyseerd met behulp van LC-MS^n en NMR en lieten verrassende en complexe chemische verbindingen zien met onder andere antimicrobiële eigenschappen. Door hierna de synthese van deze structuren in de gastheer en de bijbehorende deletiestammen te koppelen aan de uitgevoerde genetische modifcatie is een eerste stap gezet in de opheldering van hun complexe biosynthese route.

In **Hoofdstuk 3** identificeren we de structuren van vijf, in lage concentraties aanwezige, producten welke gevonden zijn in de vergelijkende metaboliet profileringsmethode van de deletiestammen zoals beschreven in **Hoofstuk 2**. Gebaseerd op deze nieuwe structuren konden de missende stappen in de biosynthese opgevuld worden. Hier zagen we dat de betrokken enzymen niet specifiek zijn maar meerdere reacties kunnen katalyseren. Dit resulteerde in de vorming van natuurlijke producten met ongebruikelijke chemische structuren welke cytotoxische en potentieel anti-kanker gerelateerde eigenschappen hebben. Uitgaande van histidyltryptophanyldiketopiperazine (HTD), dat gesynthetiseerd wordt uit tryptophan en histidine door het primaire enzym synthetase RoqA, wordt uiteindelijk roquefortine D gevormd door het enzym RoqD. Tegelijkertijd oxideert RoqR (een cytochroom p450 oxido-reductase) de histidinyl groep van HTD, waarbij dehydrohistidyltryptophanyldiketopiperazine (DHTD) wordt gevormd. Beide gelijktijdige reacties van HTD leiden tot een vertakking van de roquefortine/meleagrin route. Hierbij leidt één tak naar DHTD, via de oxidatie door RoqR, en verder naar roquefortine C door de toevoeging van een dimethylallyl groep door RoqD. In de tweede route is de dimethylallyl groep eerst toegevoegd door RoqD om roquefortine D te vormen, dat hierna wordt geoxideerd tot roquefortine C door RoqR. RoqM bestaat uit een flavine-afhankelijke MAK 1-mono-oxygenase gelijkend eiwit en is betrokken bij de omzetting van roquefortine C naar roquefortine L, een nieuwe verbinding met een ongebruikelijke nitron groep en een mogelijke hydroxyroquefortine-achtige structuur als intermediair. Roquefortine F, een component voorheen alleen beschreven in *Penicillium* soorten uit de diepe oceaan, wordt gesynthetiseerd door een opeenvolging van enzymatische modificatie door RoqM en RoqN op roquefortine C, wat uiteindelijk leidt tot de vorming van neoxaline door RoqO. Bovendien zijn RoqN en RoqO ook betrokken bij

de biosynthese van glandicoline B en meleagrin; alleen werken de enzymen dan in omgekeerde volgorde. Kortom, deze resultaten demonstreren de geschiktheid van de ontwikkelde "pipeline", voor de ontdekking en de-novo identificatie van nieuwe metabolieten, in complexe biologische monsters.

In **Hoofdstuk 4** hebben we de metaboliet ontdekking en identificatie "pipeline" succesvol toegepast bij het onderzoek naar de structuur van een cryptisch niet-linear tetra-modular NRPS in *Penicillium chrysogenum*. Hier hebben we verschillende nieuwe cyclische en lineaire tetrapeptides (een peptide bestaande uit vier aminozuren) met gelijke chemische eigenschappen, geïdentificeerd door gebruik te maken van een deletie in een enkel gen in combinatie met de vergelijkende metaboliet profileringsmethode. Doordat gebruik gemaakt is van de sequenties van de gevormde cyclische producten werd een zekere mate van aspecificiteit van alle betrokken adenylatie-domeinen in relatie tot hun substraten herleid, waardoor deopname van verschillende aminozuren in het eindproduct kan worden verklaard. Door gebruik te maken van substraat voorspelling is er voor elk molecuul een gedetailleerde synthese route herleid. Deze studie laat zien dat een uitgebreide analytische methode noodzakelijk is om meer inzicht in het metaboloom krijgen. Aangezien het mechanisme van de biosynthese is afgeleid van de aminozuursequentie van de producten, is er volledige kwantitatieve en structurele informatie van de natuurlijke producten nodig om elk adenylatie-domein te kunnen verbinden met de individuele ketenverlengingstap.

Echter, *de-novo* structuuropheldering met MS zoals is toegepast in **Hoofdstuk 2-4** is een uitdagende taak die, ondanks MS/MS fragmentatie, tot op heden nog beperkte structurele informatie oplevert. Hiernaast heeft structuuropheldering met behulp van NMR een hoge zuiverheid van het monster nodig. Om deze problemen op te lossen hebben we de eerste stappen gezet naar een geautomatiseerde *de-novo* structuuropheldering. In **Hoofdstuk 5** beschrijven we de structuurophelderings "pipeline" CMCP (chemoinformatics supported MS^n comparison pipeline). Deze CMCP is gebaseerd op de vergelijking van gegevens uit opeenvolgende fragmentatie massaspectrometrie experimenten (MS^n). Na isolatie van de metaboliet uit het biologisch monster, wordt van deze opgezuiverde fractie een serie opeenvolgende hoge-resolutie fragmentatiespectra opgenomen. Vanuit deze spectra worden fragmentatie-en 'neutral loss' bomen gegenereerd en vergeleken met een database die gevuld is met fragmentatie- en 'neutral loss' bomen van bekende verbindingen. Middels software kunnen overeenkomstige fragmentaties of 'neutral losses' worden herkend. Deze overeenkomstige fragmentatie, gebaseerd op de bekende structuur uit de database, maakt het mogelijk om zo de structuren van onbekende componenten met grote zekerheid te herleiden. Dit werkt ook zonder dat de verbinding daadwerkelijk aanwezig is in de database. Afhankelijk van de database query resultaten van de onbekende verbindingen, kunnen zowel specifieke fragmenten of 'neutral losses' identiek zijn, of beiden. Om aan te tonen dat er verschillende conceptuele uitkomsten van de database query mogelijk zijn (vergelijkingen gebaseerd op een identiek fragment en "neutral-loss" boom) en om te laten zien hoe deze informatie gebruikt kan worden, is de-novo structuuridentificatie van de eerder gevonden onbekende secundaire metabolieten roquefortine C, dehydrohistidyltryptophanyldiketopipera-

zine (DHTD), roquefortine F en roquefortine D succesvol uitgevoerd. Deze voorlopig geïdentificeerde (sub)structuren zijn uiteindelijk bevestigd met NMR.

De traditionele benaderingen voor het opsporen van metabolieten, met behulp van massaspectrometrie, gebruiken vloeistof chromatografie (LC) voorafgaand aan de metaboliet detectie, zoals is beschreven in **Hoofdstuk 2-4**. Door gebruik te maken van deze chromatografie-stap zijn er arbeidsintensieve en tijdrovende monster-voorbereidingsstappen nodig, evenals significant lange analysetijden. Bovendien is de scheiding en detectie van de metabolieten sterk afhankelijk van de chemische aard van de te bepalen stoffen en de gebruikte chromatografische methode. Daarentegen is directe infusie massaspectrometrie (DI-MS) een snelle, betrouwbare, gevoelige en kosteneffectieve methode die bijzonder aantrekkelijk is als het gaat om grote aantallen monsters. Echter, structuuridentificatie gebruik makend van directe infusie, is nog steeds moeizaam, zo niet onmogelijk aangezien co-fragmentatie van isobaren en isomeren deze toepassing beperken tot het enkel creëren van een vingerafdruk van een enkele metaboliet. In **Hoofdstuk 6** beschrijven we de ontwikkeling en toepassing van CMCP in combinatie met directe infusie voor de structuuropheldering van metabolieten. Vergelijkbaar met de identificatie van metabolieten zoals beschreven in **Hoofdstuk 5**, zijn de moleculaire structuren van onbekende metabolieten afgeleid van overeenkomstige fragmentatiemechanismen van gelijksoortig fragmenterende verbindingen in de database. Echter, in tegenstelling tot de structuuropheldering, zoals beschreven in **hoofdstuk 5**, waarbij een gezuiverde fractie van het monster werd gebruikt voor het opnemen van de fragmentatieboom, kunnen bij directe infusie ongewenste ionen, zoals isobaren en isomeren, de structuuropheldering belemmeren omdat er geen gebruik wordt gemaakt van de chromatografische voor-scheiding. Deze isobaren en isomeren zullen gelijktijdig worden gefragmenteerd met de te identificeren verbinding waardoor het MS^2 fragmentatie spectrum extra fragmentionen bevat die de structuuropheldering belemmeren. Om dit te voorkomen hebben we deze fragmentionen verder gefragmenteerd tot 3-de generatie fragmenten. De op deze wijze verkregen substructuren zijn idealiter afkomstig uit slechts één populatie van structureel identieke ionen. Dit laten we zien we voor de identificatie van roquefortine C, roquefortine F en glandicoline B. Door met CMCP naar overeenkomende fragmenten en “neutral-losses” in de verschillende fragmentatiespectra te zoeken, konden de gelijke substructuren van de in de database aanwezige verbindingen en structuren herkend worden. Aan de hand van deze identieke (sub)structuren is de structuur van de onbekende verbinding herleid. Dit was zelfs mogelijk als deze verbinding niet in de database aanwezig was. Door de integratie van directe infusie en de metaboliet ontdekking en identificatie “pipeline” wordt een analytisch platform gecreerd waarbij naast screening ook identificatie van onbekende structuren kan worden uitgevoerd.

Tot slot, laten we in dit proefschrift de ontwikkeling en toepassing van analytische platformen zien voor de ontdekking, relatieve kwantificering en identificatie van secundaire metabolieten in *Penicillium chrysogenum*. Alhoewel we dit alleen hebben toegepast in schimmelculturen kunnen deze platformen eenvoudig worden toegepast voor de ontdekking en identificatie van metabolieten in andere biologische monsters.

Vooral het gebruik van de geautomatiseerde fragmentatieboom vergelijking heeft laten zien dat dit een erg krachtige methode is voor de structuuridentificatie van metabolieten. Als we dit vergelijken met de traditionele toegepaste MS/MS benadering waarbij fragment ionen worden gebruikt in slechts één spectrum geeft de MSn benadering meerdere lagen van fragment ionen en meer inzicht in de precursor-fragment ion relatie. Omdat de structuur van de precursor is verdeeld in fragmenten geven deze subfragmentatiebomen structurele informatie over de precursors. Het vergelijk van de fragmenten in MS^n is niet gelimiteerd tot de massa-over-lading ratio's, elementaire samenstellingen of intensiteiten en laat een vergelijking van fragment structuren zien welke vele voordelen biedt ten opzichte van de MS/MS benadering. Naast het duidelijke voordeel voor de-novo identificaties zoals is beschreven in **Hoofdstuk 5** en **6** geeft de meerdere lagen fragmentatieboom vergelijking meer betrouwbaarheid in identiteitsonderzoeken van verbindingen, omdat de fragmenten en "neutral losses" vergeleken over meerdere fragmentatie lagen meer informatie bieden dan een enkele fragmentatie. Daarom denken wij dat MS^n methoden in de nabije toekomst meer en meer de MS/MS methoden zullen vervangen in onderzoeken waar gedetailleerde informatie van een metaboliet nodig is. Echter door de langere meettijden van de MS^n experimenten vergeleken met de MS/MS kunnen de complete fragmentatiebomen niet altijd gemaakt worden gedurende de metaboliet profiel metingen met de huidige generatie massaspectrometers. Daarom zijn de methoden die "on the fly" directe fractie collectie gecombineerd met directe infusie massaspectrometrie zoals gedaan kan worden met de Advion Nanomate en beschreven in **Hoofstuk 6** belangrijke methoden voor het efficiënt bepalen van een complete fragmentatieboom.

Referenties

Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D., and Keller, N.P. (2006). Genomic mining for Aspergillus natural products. Chemistry & biology 13, 31-37.

Newman, D.J., and Cragg, G.M. (2007). Natural products as sources of new drugs over the last 25 years. J Nat Prod 70, 461-477.

Zerikly, M., and Challis, G.L. (2009). Strategies for the discovery of new natural products by genome mining. Chembiochem 10, 625-633.

Appendix

Acknowledgements
Curriculum Vitae
List of publications

Curriculum Vitae

Marco Ries was born on October 16th, 1982. After his high-school diploma in 2002 and 9 months of military service, Marco studied chemistry at the Johannes-Gutenberg University (Mainz, Germany) and University of Valencia (Valencia, Spain). During that time, he worked for Clariant GmbH (Frankfurt, Germany) and Sanofi-Aventis GmbH (Frankfurt, Germany), among others.
After graduating with a diploma thesis in the field of analytical environmental chemistry under the supervision of Prof. Hoffmann in Mainz, Marco started his PhD project at the division of Analytical BioSciences of the University Leiden (Leiden, Netherlands) in November 2009.
Since October 2013, Marco is writing his diploma thesis at the department for quantitative methods and mathematical economics of the University Hagen (Hagen, Germany) to complete his diplomas in business and economics.

List of publications

Application of time-of-flight aerosol mass spectrometry for the online measurement of gaseous molecular iodine
Michael Kundel, Ru-Jin Huang, Ute Thorenz, Janine Bosle, Moritz Mann, Marco Ries, Thorsten Hoffmann
Anal Chem. 2012; 84 (3): 1439-1445

A branched biosynthetic pathway is involved in production of roquefortine and related compounds in Penicillium chrysogenum
Hazrat Ali, Marco Ries*, Jeroen Nijland, Peter Lankhorst, Thomas Hankemeier, Roel Bovenberg, Rob Vreeken , Arnold Driessen*
PLoS One. 2013; 8 (6): e65328

Novel key metabolites reveal further branching of the roquefortine/meleagrin biosynthetic pathway
Marco Ries, Hazrat Ali*, Peter Lankhorst, Thomas Hankemeier, Roel Bovenberg, Arnold Driessen, Rob Vreeken*
J. Biol. Chem. 2013; 288 (52): 37289-37295

A single unspecific non-linear NRPS is involved in the synthesis of cyclic tetrapeptides in Penicillium chrysogenum
Hazrat Ali, Marco Ries*, Peter Lankhorst, Rob van der Hoeven, Olaf Schouten, Marek Noga, Thomas* Hankemeier, *Noel van Peij, Roel Bovenberg, Rob Vreeken, Arnold Driessen*
Submitted to PLoS One

Chemoinformatics supported MS^n Comparison Pipeline (CMCP): Towards automated *de novo* structure elucidation using multiple-stage fragmentation tree comparison
Marco Ries, Jeroen Kazius, Hazrat Ali, Arnold Driessen, Thomas Hankemeier, Theo Reijmers, Rob Vreeken
Manuscript in preparation

Multiple stage fragmentation tree comparison enables detailed structure elucidation in direct infusion mass spectrometry based experiments
Marco Ries, Hazrat Ali, Arnold J.M. Driessen, Thomas Hankemeier, Theo Reijmers, Rob J. Vreeken
Manuscript in preparation

* Authors contributed equally to this work